

# La struttura dell'impossibile

*Considerazioni su una teoria unificata dei limiti  
del pensiero, della vita e del calcolo*

*Seconda edizione riveduta e ampliata*

# Nota Metodologica Preliminare

## I. Tre livelli di claim: una bussola epistemica per il lettore

Un'opera che si propone di unificare impossibilità strutturali provenienti da campi eterogenei — logica matematica, biologia evolutiva, teoria della complessità computazionale, filosofia della mente — si espone a un rischio specifico: quello di presentare come dimostrazioni ciò che è congettura, o come congettura ciò che è semplice proposta interpretativa. Una buona narrazione può simulare rigore senza possederlo.

Per rendere questo ciclo intellettualmente onesto, ogni capitolo è preceduto da una 'mappa epistemica' che classifica i claim principali in tre categorie. I marcatori [D], [C], [I] compaiono nei riquadri colorati nel testo.

| Cod.     | Nome                    | Definizione                                                                                                                   | Come riconoscerlo                                                                                              |
|----------|-------------------------|-------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| <b>D</b> | Formalmente Dimostrato  | Catena completa di passi deduttivi dagli assiomi. Non dipende da congetture aperte.                                           | Gödel, Tarski, Turing, JL, Landauer. Ogni dipendenza da $P \neq NP$ è sempre esplicitata.                      |
| <b>C</b> | Congettura Fondata      | Evidenza solida, accettato dalla comunità, ma senza dimostrazione formale completa o dipendente da assunzioni non dimostrate. | $P \neq NP$ ; il PAI nella forma generale; la struttura $\mathbb{Z}$ ; la legge di conservazione della carica. |
| <b>I</b> | Proposta Interpretativa | Analogia strutturale illuminante tra un risultato noto e il confine sintassi-semantica. Non è una riduzione formale.          | Paradosso di Rogers; divario esplicativo; problema della misurazione quantistica; choice blindness.            |

## II. Una distinzione strutturale, non retorica

La distinzione D/C/I non è un gesto di modestia. È una distinzione con conseguenze operative precise. Un risultato D è assoluto rispetto al sistema assiomatico: non è messo in discussione da nessuna osservazione empirica o argomentazione filosofica. Può soltanto essere relativizzato (dimostrato vero solo dentro certi assiomi) o esteso. Il teorema di Gödel non è contestato da nessun risultato delle neuroscienze cognitive.

Un risultato  $C$  è revocabile. Se un giorno qualcuno dimostrasse  $P = NP$ , tutti i claim  $C$  che dipendono da  $P \neq NP$  collasserebbero. Un risultato  $I$  è più fragile e più ricco: fragile perché dipende dalla fertilità dell'analogia, ricco perché è la fonte principale della creatività teorica. Le grandi sintesi iniziano quasi sempre come proposte interpretative prima di diventare, eventualmente, congetture o dimostrazioni.

### III. Chiarimento sul PAI e sulla struttura $\mathbb{Z}_5$

#### [CONGETTURA FONDATA]

Il Principio di Auto-Rappresentazione Insufficiente (PAI, Capitolo VII) non è un teorema matematico nel senso dei teoremi di Gödel, Tarski e Turing. È un 'teorema filosofico con dimostrazione condizionale': la sua parte formale (le condizioni  $C1$  e  $C2$  seguono dal teorema di Tarski) è genuinamente dimostrata; la parte che separa  $C3$  da  $C4$  dipende da  $P \neq NP$ . Il PAI nella sua forma a quattro condizioni ha quindi lo stesso status epistemico di  $P \neq NP$ : quasi certamente vero, non formalmente dimostrato.

Allo stesso modo, la struttura  $\mathbb{Z}_5$  delle implicazioni reciproche tra le cinque famiglie (Capitolo XII) è una congettura strutturale fondata, non un teorema dimostrato. Le implicazioni  $F_1 \rightarrow F_2$  (via Brouwer) e  $F_4 \rightarrow F_5$  (via Landauer) sono dimostrate; le implicazioni  $F_2 \rightarrow F_3$ ,  $F_3 \rightarrow F_4$ ,  $F_5 \rightarrow F_1$  dipendono da  $P \neq NP$  o dalla condizione di auto-rappresentazione.

### IV. Nota sul linguaggio

Il termine 'paesaggio' (nei capitoli II, V, VI e XII) è usato nel senso tecnico della biologia evolutiva e dell'ottimizzazione: uno spazio multidimensionale in cui ogni punto corrisponde a una configurazione del sistema e la sua altitudine a una funzione di valore. In alcune sezioni della versione precedente il termine era stato erroneamente tipografato come 'paecapitolo' per errore di composizione: questa versione usa 'paesaggio' in modo coerente.

Il termine 'sistema auto-rappresentazionale' è usato nel senso formale della coppia  $(S, \varphi)$  dove  $S$  è un sistema formale e  $\varphi$  è una funzione di auto-applicazione, come definito nel Capitolo VII.



# Capitolo I

## Il catalogo e la tassonomia delle impossibilità

### Verso una classificazione formale dei limiti strutturali

#### Mapa epistemica — Capitolo I

| Status          | Cod. | Claim                                                                                                                                                                                      |
|-----------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | La distinzione tra impossibilità contingente e strutturale si formalizza attraverso la nozione di proprietà universale in teoria delle categorie.                                          |
| DIMOSTRATO      | D    | Le quattro famiglie di impossibilità hanno strutture matematiche distinte e non sono riducibili le une alle altre in modo elementare.                                                      |
| CONGETTURA      | C    | La tesi centrale — tutte le impossibilità come manifestazioni del confine sintassi-semantica — è una proposta interpretativa nella sua versione forte, una congettura strutturale fondata. |
| INTERPRETAZIONE | I    | La Famiglia IV (chiusura semantica) come la famiglia ‘più profonda’ di cui le altre sono manifestazioni: argomento illuminante, non riduzione formale.                                     |

### I. Il problema che i due cicli non hanno posto

I due cicli precedenti — Ai confini del calcolabile e Ai confini del vivente — hanno accumulato un nutrito catalogo di risultati negativi: teoremi di incompletezza, problemi indecidibili, limiti della rappresentazione neurale, impossibilità del binding esatto, divario esplicativo, ratchet di Müller, paradosso di Rogers, teorema No Free Lunch, soglia demografica di Henrich, chiusura dei periodi critici, limiti dell'inferenza causale. Ciascuno di questi risultati è stato sviluppato nel proprio contesto disciplinare, con la propria strumentazione tecnica, con i propri antenati intellettuali e le proprie conseguenze applicative.

Ma i due cicli non hanno posto il quesito che la loro giustapposizione rende inevitabile: questi risultati negativi sono davvero eterogenei — difetti accidentali di specifici sistemi, manifestazioni della complessità di specifici domini — oppure condividono una struttura profonda che li lega in qualcosa di più unificato? La risposta che questo capitolo inaugura, e che i successivi sviluppano con sufficiente rigore, è che non sono eterogenei. Appartengono a un insieme di tipi formalmente distinti, ciascuno dei quali ha la propria struttura matematica, le proprie condizioni

di emergenza e le proprie conseguenze epistemologiche. La tassonomia di questi tipi è il contributo principale del presente contributo.

Prima di costruire la tassonomia, però, occorre chiarire cosa si intende per "impossibilità strutturale" — in modo da escludere dall'analisi i casi che non meritano questo nome.

## **II. La distinzione tra impossibilità contingente e strutturale**

Non ogni limitazione è una impossibilità strutturale nel senso rilevante per questo ciclo. Si possono distinguere almeno quattro gradi di limitazione che è necessario tenere separati con precisione.

Il primo grado è la limitazione empirica contingente: la limitazione che dipende da fattori accidentali dello stato attuale della conoscenza o della tecnologia. Il fatto che nel 1900 non esistesse un vaccino contro la poliomielite era una limitazione del secondo tipo — contingente: la poliomielite era biologicamente guaribile, e il vaccino esisteva nello spazio delle possibilità biologiche, non era ancora stato trovato. Questo tipo di limitazione non interessa questo ciclo.

Il secondo grado è la limitazione pratica: la limitazione che dipende da vincoli di risorse (tempo, energia, materia) ma non da impossibilità logiche o matematiche. Il calcolo dell'enumerazione esplicita di tutti i genotipi possibili di un genoma umano ( $2^{(2L)}$  con  $L = 10^6$  loci) è praticamente impossibile per vincoli di complessità computazionale — nessun computer fisicamente realizzabile potrebbe completarlo nell'arco della vita dell'universo — ma non è logicamente impossibile: un agente con risorse infinite potrebbe farlo. Questo tipo è più interessante del primo ma rimane esterno al tema principale.

Il terzo grado è la limitazione formale interna a un sistema: la limitazione che dipende dalla struttura assiomatica di un sistema formale specifico e potrebbe in principio essere superata cambiando il sistema. Le proposizioni indecidibili di Peano sono indecidibili in PA ma dimostrabili in sistemi più forti. Queste limitazioni sono estremamente interessanti ma non costituiscono il nucleo di questo ciclo: sono contingenti rispetto alla scelta del sistema.

Il quarto grado — il solo che questo ciclo tratta sotto il nome di "impossibilità strutturale" — è la limitazione che emerge da proprietà necessarie di qualsiasi

sistema sufficientemente ricco per porre la domanda in questione. L'incompletezza di Gödel non è una proprietà peculiare di PA: è una proprietà di qualsiasi sistema formale consistente che contenga abbastanza aritmetica per codificare la propria sintassi. Il divario esplicativo non è una lacuna della neuroscienza attuale: è la conseguenza strutturale del fatto che le descrizioni fisiche e le descrizioni fenomeniche usano tipi di concetti che non hanno implicazioni a priori reciproche. Il binding esatto è impossibile non per mancanza di neuroni ma per il teorema di dimensionalità che abbiamo dimostrato nel capitolo I del ciclo biologico.

La formalizzazione della distinzione usa la nozione di proprietà universale in teoria delle categorie. Un'impossibilità è strutturale se è una proprietà di tutti gli oggetti in una certa categoria — non di qualche oggetto specifico. Formalmente: un enunciato  $E$  esprime una impossibilità strutturale rispetto alla categoria  $C$  se

$$\forall O \in \text{Ob}(C): E(O)$$

con  $E(O) = "O \text{ non può [fare } X]"$ . L'incompletezza di Gödel è strutturale nella categoria dei sistemi formali consistenti che contenga l'aritmetica; il divario esplicativo è strutturale nella categoria delle teorie fisiche strutturaliste; il ratchet di Müller è strutturale nella categoria dei sistemi di replicazione sequenziale senza ricombinazione.

### III. Il censimento: quattro famiglie di impossibilità

Dopo avere applicato il criterio di strutturalità al catalogo dei risultati negativi accumulati nei due cicli precedenti, emergono quattro famiglie distinte — quattro tipi di impossibilità con strutture matematiche diverse, anche se, come si vedrà nei capitoli successive, non del tutto indipendenti.

#### *Famiglia I: Le impossibilità diagonali*

Questa famiglia comprende tutti i risultati che si dimostrano attraverso una variante della costruzione diagonale — la tecnica inventata da Cantor nel 1891, applicata da Russell nel 1902, da Gödel nel 1931, da Turing nel 1936, da Kleene e da tutti i successivi teoremi di indecidibilità. La struttura è sempre la stessa: si assume l'esistenza di un oggetto  $X$  con certe proprietà di completezza o decidibilità, si costruisce un nuovo oggetto  $Y$  definito in termini di  $X$  attraverso

un'operazione di auto-riferimento, e si mostra che  $Y$  porta a contraddizione o alla confutazione dell'assunzione iniziale.

La forma canonica della costruzione diagonale è la seguente. Sia  $f: A \rightarrow B$  una funzione che si suppone suriettiva (o totalmente calcolabile, o completa nel senso di coprire tutti i casi rilevanti). Si definisce un elemento  $d \in B$  attraverso la regola diagonale: per ogni  $a \in A$ ,  $d$  "differisce" da  $f(a)$  in qualche dimensione rilevante (si nega il valore, si nega la terminazione, si afferma l'indimostrabilità). Si mostra che  $d$  non può essere nell'immagine di  $f$  (o nell'estensione della funzione, o nel dominio di verifica), contraddicendo l'assunzione iniziale.

Il teorema di Cantor è il prototipo: per qualsiasi funzione  $f: \mathbb{N} \rightarrow \{0,1\}^{\mathbb{N}}$ , la sequenza diagonale  $d$  definita da  $d_n = 1 - f(n)_n$  non è nell'immagine di  $f$ . Il teorema di Gödel è una variante più sofisticata: la sentenza  $G = "G \text{ non è dimostrabile in PA}"$  è costruita attraverso la codifica gödeliana che permette di definire  $d$  in termini di  $f$  (la funzione di dimostrabilità di PA), e  $G$  non può essere né dimostrata né refutata in PA. Il problema dell'arresto di Turing è un'altra variante: il programma diagonale  $D$ , definito come " $D(x)$ : se  $H(x,x) = \text{HALT}$  allora LOOP, altrimenti HALT", contraddice qualsiasi supposta macchina di Turing  $H$  che decida l'arresto.

nei capitoli dei due cicli precedenti, le impossibilità diagonali compaiono esplicitamente nel teorema di Minsky-Papert (la costruzione della sequenza non separabile linearmente), nel teorema di Rice (la diagonalizzazione sul comportamento semantico dei programmi), e implicitamente nel problema del frame (la costruzione di una proprietà di rilevanza auto-referenzialmente indefinita). Il capitolo II di questo ciclo sviluppa questa famiglia con la massima profondità.

### *Famiglia II: Le impossibilità dimensionali*

Questa famiglia comprende i risultati che emergono da una discrepanza tra la dimensionalità delle risorse disponibili e la dimensionalità del problema da risolvere. Non si tratta di limitazioni di complessità computazionale in senso stretto (che appartengono alla Famiglia III), bensì di impossibilità geometriche: lo spazio delle soluzioni è topologicamente più ricco dello spazio delle rappresentazioni disponibili.

Il teorema di impossibilità del binding esatto (Capitolo I del ciclo biologico) è l'esempio paradigmatico: per  $n$  neuroni, non è possibile rappresentare senza errori più di  $n$  feature simultaneamente in uno spazio di dimensione  $n$ , e qualsiasi soluzione approssimata introduce interferenze di ampiezza  $\sqrt{(n/m)(1-n/m)}$  per  $m > n$  feature. La struttura geometrica è quella del teorema di Johnson-Lindenstrauss nella sua forma converse: la proiezione da uno spazio di dimensione  $m$  in uno spazio di dimensione  $n < m$  non può preservare tutte le distanze — qualche coppia di punti deve essere compressa.

La formula di Cover per la separabilità lineare appartiene alla stessa famiglia:  $C(m, n) \rightarrow 0$  per  $m > 2n$  mostra che la capacità di separazione lineare si esaurisce quando il numero di pattern supera il doppio della dimensione dello spazio. Il teorema di Minsky-Papert sul predicato di parità è una sua istanza combinatoriale: la parità su  $n$  variabili richiede accesso globale a tutte le variabili (dimensione  $n$ ) e non è approssimabile da classificatori con connettività sub-lineare (dimensione  $< n$ ).

Nella biologia, la soglia demografica di Henrich appartiene a questa famiglia: la complessità sostenibile di una cultura è limitata dalla dimensione della rete di apprendimento — una limitazione dimensionale nel senso che il numero di "parametri" della tecnica culturale deve stare dentro la "capacità" della rete. La formula  $n^* = \exp((\theta_{\max}/\sigma)^2/2)$  è la versione demografica del teorema di copertura.

Caratteristica formale distintiva della Famiglia II: le impossibilità dimensionali sono necessariamente quantitative — non dicono soltanto che qualcosa è impossibile, ma specificano la funzione di scambio tra dimensionalità e fedeltà della rappresentazione. Questa precisione quantitativa le rende particolarmente sfruttabili come strumenti ingegneristici: sapendo la funzione di scambio, si può calcolare il costo di ogni approssimazione.

### *Famiglia III: Le impossibilità di complessità*

Questa famiglia comprende i risultati che derivano dalla complessità computazionale intrinseca di certi problemi — la tesi che certi problemi non abbiano soluzione in tempo polinomiale (assumendo  $P \neq NP$ ), e più in generale che appartengano a classi di complessità che eccedono le risorse disponibili. La

distinzione dalla Famiglia I è sottile ma precisa: le impossibilità diagonali mostrano che certi problemi non hanno soluzione in linea di principio da nessun algoritmo; le impossibilità di complessità mostrano che certi problemi non hanno soluzione efficiente — esistono algoritmi che li risolvono, ma richiedono risorse superpolinomiali.

La  $\Pi_2^P$ -completezza della Circumscription (teorema di Liberatore e Schaerf, discusso nel capitolo III del ciclo biologico) appartiene a questa famiglia: non è che il problema del frame sia in linea di principio irrisolvibile — è che qualsiasi sua soluzione generale richiede risorse computazionali che eccedono qualsiasi sistema fisicamente realizzabile in tempo reale. La NP-hardness del rilevamento delle comunità nei grafi (discussa nel capitolo X del ciclo biologico per la massimizzazione della modularità Q) è un altro caso: l'algoritmo ottimale esiste ma richiede tempo esponenziale.

Le impossibilità di complessità hanno una struttura formale molto diversa dalle impossibilità diagonali: dipendono da una congettura non dimostrata ( $P \neq NP$ ), anche se la congettura è considerata quasi certamente vera dalla comunità matematica. Questo le rende epistemologicamente più fragili: in teoria, una dimostrazione di  $P = NP$  eliminerebbe questa intera famiglia di impossibilità. Ma nella pratica, la solidità della congettura è tale che le impossibilità di complessità sono trattate come impossibilità a tutti gli effetti operativi.

#### *Famiglia IV: Le impossibilità di chiusura semantica*

Questa famiglia — la più originale dell'analisi proposta in questo ciclo — comprende i risultati che emergono dalla impossibilità per un sistema formale o fisico di rappresentare o implicare la propria semantica dall'interno. La caratteristica distintiva è che il sistema può descrivere perfettamente la propria struttura sintattica ma non può catturare il proprio significato.

Il teorema di Tarski sull'indefinibilità della verità è il caso prototipico: nessun sistema formale consistente e sufficientemente ricco può definire al proprio interno il predicato di verità per il proprio linguaggio. Qualsiasi definizione di "verità in L" deve essere espressa in un metalinguaggio L' più ricco di L — e lo stesso argomento si applica a L', richiedendo un meta-metalinguaggio L'', e così via

nella gerarchia tarskiana. Il sistema non può mai "chiudersi" sulla propria semantica.

Il divario esplicativo di Chalmers appartiene a questa famiglia: le descrizioni fisiche sono descrizioni sintattiche (descrivono relazioni strutturali tra entità), e i concetti fenomenici sono concetti semantici (si riferiscono a proprietà intrinseche dell'esperienza che non sono relazionali). L'impossibilità di derivare i secondi dai primi non è una limitazione della fisica attuale — è la conseguenza strutturale del fatto che concetti di tipo sintattico non implicano a priori concetti di tipo semantico. Il "carattere fisico" e il "carattere fenomenico" appartengono a livelli della gerarchia di Tarski che non comunicano direttamente.

La separazione wanting/liking di Berridge (Capitolo XIV del ciclo biologico) è una versione biologica di questa impossibilità: il sistema dopaminergico (che implementa la sintassi dell'incentivazione) e il sistema oppioide (che implementa la semantica del piacere) sono dissociabili proprio perché operano su livelli diversi della gerarchia semantica. Il "volere qualcosa" e il "godere di qualcosa" non sono aspetti dello stesso processo — sono processi che appartengono a livelli diversi della struttura causale del valore.

Il paradosso di Rogers (Capitolo VII del ciclo biologico) è un'altra istanza: l'apprendimento sociale puro non aumenta la fitness media della popolazione perché copia la struttura sintattica dei comportamenti (i comportamenti osservabili) senza accedere alla struttura semantica che li ha prodotti (l'adattamento all'ambiente specifico). La copia del comportamento senza copia delle ragioni del comportamento è una trasmissione di sintassi senza semantica — e la semantica è precisamente ciò che ha valore adattivo.

#### **IV. La struttura delle famiglie: separazione e intersezione**

Le quattro famiglie non sono compartimenti stagni. Esistono risultati che appartengono genuinamente a più di una famiglia, e la loro intersezione rivela connessioni strutturali che una trattazione puramente disciplinare non avrebbe potuto scoprire.

Il teorema di incompletezza di Gödel appartiene sia alla Famiglia I (è dimostrato per diagonalizzazione) sia alla Famiglia IV (afferma l'esistenza di verità che

trascendono il sistema — un'impossibilità di chiusura semantica): la sentenza G è indimostrabile perché il sistema non può "vedere" la propria semantica dall'interno. La diagonalizzazione è il meccanismo della dimostrazione; la chiusura semantica è ciò che la diagonalizzazione rivela essere impossibile.

Il problema del frame appartiene sia alla Famiglia III (è  $\Pi_2^P$ -completo) sia alla Famiglia IV (il concetto di "rilevanza" è semantico e non può essere ridotto a un criterio sintattico): anche se il calcolo fosse infinitamente potente, non potrebbe risolvere il frame problem perché la rilevanza non è una proprietà formalmente definibile — è una proprietà che dipende dall'interpretazione dei simboli nel contesto, non dalla struttura dei simboli stessi.

Il ratchet di Müller appartiene alla Famiglia II (è una limitazione della capacità di correzione degli errori in funzione della dimensione della popolazione e del tasso di mutazione) ma ha una struttura che verrà analizzata nel capitolo VI come caso di impossibilità irreversibile — una quinta famiglia che emerge dall'analisi degli aspetti dinamici dei limiti strutturali e che per ora lasciamo come promessa di approfondimento.

La struttura delle intersezioni può essere rappresentata formalmente attraverso la nozione di riduzione tra classi di impossibilità. La Famiglia I si riduce alla Famiglia IV nel seguente senso preciso: le impossibilità diagonali sono, nella loro forma più profonda, impossibilità di auto-referenza semantica — sono la dimostrazione che un sistema non può descrivere la propria semantica dall'interno. La riduzione formale è questa: ogni costruzione diagonale può essere vista come la costruzione di un oggetto d che "dice qualcosa su se stesso" (auto-riferimento semantico) — e l'impossibilità emerge proprio dal fatto che il sistema deve usare la propria struttura sintattica per fare affermazioni semantiche su quella struttura, il che genera circolarità irrisolvibile.

Questo suggerisce che la Famiglia IV non è una tra quattro famiglie equivalenti: è la famiglia più profonda, quella di cui le altre sono manifestazioni specializzate. il capitolo III svilupperà questa tesi con la precisione che merita, mostrando che il confine sintassi-semantica è il locus generativo di tutte le impossibilità strutturali.

## V. La misura delle impossibilità: gradi e gerarchie

Non tutte le impossibilità sono ugualmente impossibili — e questa affermazione non è solo un gioco di parole. Le impossibilità hanno struttura graduata che può essere misurata con precisione formale.

Per le impossibilità diagonali, la misura è la gerarchia aritmetica: i problemi  $\Sigma_0^0 = \Pi_0^0 = \Delta_1^0$  sono decidibili; i problemi  $\Sigma_1^0$  (semi-decidibili) includono il problema dell'arresto; i problemi  $\Pi_1^0$  includono i complementi dei problemi semi-decidibili (ad esempio, "questo programma non termina mai"); la gerarchia continua a livelli superiori attraverso  $\Sigma^0$  e  $\Pi^0$  con complessità crescente. Il teorema di Post stabilisce che questa gerarchia è strettamente crescente:  $\Sigma^0 \subsetneq \Sigma_{+1}^0$  per ogni  $n$ . Ogni livello della gerarchia è strettamente più impossibile del precedente.

Per le impossibilità di complessità, la misura è la gerarchia polinomiale:  $P \subseteq NP \subseteq PH \subseteq PSPACE \subseteq EXPTIME \subseteq \dots$ . La  $\Pi_2^P$ -completezza della Circumscription (Famiglia III del problema del frame) colloca quel problema due livelli sopra  $P$  nella gerarchia polinomiale — più difficile di  $NP$ -completo, che è già considerato intrattabile.

Per le impossibilità dimensionali, la misura è quantitativa e continua: la funzione di scambio tra dimensionalità della rappresentazione  $n$  e fedeltà della codifica (ad esempio, l'errore di interferenza  $\varphi \cdot m/n$  per il binding, o la soglia demografica  $n^* = \exp((\theta_{\max}/\sigma)^2/2)$  per la trasmissione culturale) fornisce una misura continua della distanza dall'impossibilità. Il sistema non è "impossibile" o "possibile" seccamente — è impossibile-con-errore- $\epsilon$  per ogni  $\epsilon > 0$  sotto certe condizioni.

Per le impossibilità di chiusura semantica, la misura è la gerarchia tarskiana: il sistema  $L_0$  non può definire la propria verità, ma  $L_1$  (il metalinguaggio di  $L_0$ ) può definire la verità di  $L_0$ ;  $L_1$  non può definire la propria verità, ma  $L_2$  la definisce; e così via. Il livello della gerarchia a cui il sistema si trova misura quanta auto-riflessione semantica è possibile prima di incontrare il limite di Tarski.

Questa struttura graduata ha una conseguenza che la filosofia dei limiti ha spesso trascurato: le impossibilità non sono semplicemente muri che fermano il pensiero, ma topografie che il pensiero può esplorare. Sapere dove si trova il limite, e di

quale tipo di limite si tratta, permette di navigarne il perimetro — di avvicinarsi asintoticamente senza aspettarsi di attraversarlo.

## **VI. L'asimmetria delle impossibilità: perché alcuni sistemi ci arrivano e altri no**

Una delle osservazioni più significative che emergono dal confronto dei due cicli precedenti è un'asimmetria notevole: i sistemi biologici — cervelli, ecosistemi, organismi — non "scoprirono" i propri limiti dall'esterno attraverso un'analisi logica. Vi si imbattono dall'interno attraverso l'evoluzione, e spesso trovarono soluzioni parziali che abitano il perimetro del limite senza attraversarlo. I sistemi artificiali — programmi, algoritmi, computer — trovano i propri limiti esattamente dove la teoria predice, senza nessuna eccezione.

Questa asimmetria non è accidentale. I sistemi biologici sono stati selezionati per funzionare bene in regimi specifici, il che significa che si sono adattati ai propri limiti strutturali in modo tale da evitarli nei casi praticamente rilevanti mentre non è possibile evitarli nei casi più generali. Il cervello non risolve il problema del frame in linea generale — ma i prior ecologici (discussi nel capitolo III del ciclo biologico) lo evitano sistematicamente nei domini evolutivamente rilevanti. L'impossibilità del binding esatto non impedisce il funzionamento cognitivo perché la sparsità dell'attivazione ( $\varphi \approx 0.01-0.05$ ) mantiene le interferenze a livelli tollerabili nella pratica quotidiana.

I sistemi artificiali non hanno questo adattamento evolutivo: li si può costruire in modo tale da incontrare i limiti strutturali con piena forza. Un programma che cerca di risolvere il problema dell'arresto in generale non si "adatta" al limite — vi si scontra direttamente. Un classificatore lineare applicato a pattern non linearmente separabili non trova soluzioni approssimate con errore tollerabile — fallisce definitivamente.

La formalizzazione di questa asimmetria usa la nozione di "nicchia del limite" — la regione del paecapitolo dei problemi in cui un sistema può funzionare senza incontrare il limite strutturale. Sia  $L$  il limite strutturale e sia  $D$  il dominio dei problemi. La nicchia del limite è

$$N_L = \{d \in D : \text{il sistema può funzionare su } d \text{ senza incontrare } L\}$$

Un sistema biologico selezionato evolutivamente occupa  $N_L$  per il suo EEA; un sistema artificiale di uso generale cerca di operare su tutto  $D$ , incluse le regioni dove  $L$  è incontrato direttamente.

Questo ha una conseguenza di ordine pratico: la comprensione della struttura dei limiti è più urgente per i sistemi artificiali che per quelli biologici. I sistemi biologici hanno già trovato — attraverso l'evoluzione — le soluzioni pratiche al problema. I sistemi artificiali devono trovare quelle soluzioni attraverso l'ingegneria deliberata, e per farlo devono capire la struttura dei limiti con la stessa profondità con cui la selezione naturale li ha capiti implicitamente nei miliardi di anni che ha impiegato per costruire il cervello.

## **VII. La posizione epistemica di questo ciclo: tra la matematica e la filosofia**

Questo ciclo è un'impresa ibrida che abita una zona di confine tra la matematica, la biologia teorica, la filosofia della mente e la teoria dell'informazione. Non è un trattato matematico nel senso stretto: non dimostra nuovi teoremi, non produce nuovi risultati tecnici nel senso delle discipline originarie. È invece un'analisi strutturale — un tentativo di identificare l'architettura nascosta che connette risultati tecnici apparentemente eterogenei.

Questa posizione epistemica ha un antecedente preciso nella letteratura: il rapporto tra la logica matematica di Gödel, Turing e Church e la filosofia della computazione di Turing, Gandy e Sieg. I risultati tecnici (incompletezza, indecidibilità, Church-Turing) erano stati dimostrati con piena rigore matematico; la loro interpretazione filosofica — cosa significano per la natura della mente, del calcolo, della verità — ha richiesto decenni di lavoro filosofico che non era riducibile alla matematica ma non poteva prescindere da essa.

Questo ciclo si propone di fare qualcosa di analogo, ma a scala più ampia: identificare l'architettura comune dei risultati di incompletezza, delle impossibilità biologiche e dei limiti della cognizione, e articolare il significato di quella architettura per la comprensione di ciò che un sistema — computazionale, biologico, cognitivo — può e non può fare in linea di principio.

La tesi centrale, che i capitoli successive sviluppano, è la seguente: tutte le impossibilità strutturali che i due cicli precedenti hanno documentato sono manifestazioni del confine sintassi-semantica — il confine tra la struttura formale di un sistema e il significato di quella struttura. Questo confine non è una proprietà accidentale di alcuni sistemi: è una proprietà necessaria di qualsiasi sistema sufficientemente ricco da esprimere auto-riferimento. E il fatto che sistemi così diversi — cervelli biologici, sistemi formali, ecologie, culture — incontrino il medesimo confine nelle proprie rispettive impossibilità strutturali non è una coincidenza: è la traccia di una struttura profonda che questo ciclo si propone di portare alla luce.

### **VIII. Un'avvertenza metodologica**

Una teoria delle impossibilità che voglia essere rigorosa deve affrontare un rischio specifico che nessun'altra impresa intellettuale affronta nella stessa misura: il rischio di confondere impossibilità strutturali genuine con la propria incapacità di trovare soluzioni. La storia del pensiero è piena di "impossibilità" che si sono rivelate contingenti — problemi che sembravano intrattabili e che nuove prospettive hanno reso risolvibili. L'impossibilità della geometria non euclidea fu sostenuta per secoli prima che Gauss, Bolyai e Lobachevsky mostrassero che era una scelta convenzionale, non una necessità logica.

Per questo, i capitoli di questo contributo distinguono sempre esplicitamente tra impossibilità dimostrate (dove esiste un argomento formale che esclude la soluzione per qualsiasi sistema della classe pertinente) e impossibilità congetturate (dove l'argomentazione è forte ma manca di una dimostrazione formale completa). Il teorema di incompletezza di Gödel è una impossibilità dimostrata; la  $P \neq NP$  congettura è una impossibilità congetturata. Il divario esplicativo è un'impossibilità argomentata filosoficamente con forza considerevole ma non dimostrata matematicamente.

La distinzione non è soltanto accademica: le impossibilità dimostrate definiscono confini assoluti attorno a cui qualsiasi ingegneria deve navigare; le impossibilità congetturate definiscono confini probabili che meritano cautela ma non precludono l'esplorazione. Un ciclo che non mantiene questa distinzione rischia di trasformare la filosofia dei limiti in una celebrazione del pessimismo intellettuale

— esattamente il contrario di ciò che una teoria dei limiti, correttamente interpretata, dovrebbe produrre.

Come si vedrà nei capitoli conclusive di questo ciclo, la comprensione dei limiti non è fonte di scoraggiamento ma di orientamento: sapere dove finisce ciò che è possibile è la condizione per esplorare quel territorio con il massimo di profondità invece di sprecare risorse in direzioni precluse. I confini non sono muri — sono mappe.

---

## Capitolo II

### L'argomento diagonale come struttura universale

La diagonalizzazione nelle sue forme note e occulte, e la sua comparsa inaspettata nelle impossibilità biologiche

#### Mappa epistemica — Capitolo II

| Status          | Cod. | Claim                                                                                                                                                                           |
|-----------------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | Il teorema di Lawvere (1969) unifica tutti i teoremi diagonali nella struttura formale delle categorie cartesiane chiuse.                                                       |
| DIMOSTRATO      | D    | Cantor, Gödel, Turing e Tarski sono istanze del teorema di Lawvere.                                                                                                             |
| CONGETTURA      | C    | La costruzione diagonale come struttura dell'auto-applicazione in qualsiasi sistema sufficientemente ricco: congettura strutturale ben fondata.                                 |
| INTERPRETAZIONE | I    | Le forme 'occulte' della diagonalizzazione nelle impossibilità biologiche (binding, No Free Lunch, paradosso di Rogers): analogie strutturali, non riduzioni formali a Lawvere. |

#### I. Un trucco o una struttura?

Quando Georg Cantor dimostrò nel 1891 che l'insieme dei numeri reali è più grande dell'insieme dei numeri naturali, usò un argomento che appariva come un tour de force tecnico — quasi un trucco. Si suppone di avere una lista di tutti i numeri reali nell'intervallo  $[0,1]$ , scritti come decimali infiniti. Si costruisce un numero  $d$  che differisce dalla prima entrata nella prima cifra decimale, dalla seconda entrata nella seconda cifra, dalla terza entrata nella terza, e così via. Questo numero  $d$  non può apparire nella lista — per qualsiasi posizione  $n$  in cui dovesse trovarsi, differisce dalla  $n$ -esima entrata nell' $n$ -esima cifra. La lista non è completa:  $|\mathbb{R}| > |\mathbb{N}|$ .

L'impressione del trucco è rafforzata dalla semplicità della costruzione. Ma negli ottant'anni successivi alla dimostrazione di Cantor, lo stesso schema riapparve in contesti sempre più distanti e sempre più profondi: nel paradosso di Russell (1902), nel teorema di incompletezza di Gödel (1931), nel problema dell'arresto di Turing (1936), nel teorema di Rice (1953), nel teorema di Tarski (1936). In ciascuno di questi casi, il "trucco" di Cantor funzionava — e la domanda naturale

era: perché? Cosa c'è nella costruzione diagonale che la rende così universalmente applicabile?

La risposta che questo capitolo propone — e che la ricerca di Yanofsky (2003), di Lawvere (1969) in teoria delle categorie, e di Smullyan (1992) nell'approccio agli operatori del punto fisso hanno sviluppato parzialmente in contesti tecnici — è che la diagonalizzazione non è un trucco ma una struttura: è la struttura dell'auto-applicazione, la struttura di un sistema che rivolge su se stesso gli strumenti che usa per descrivere il mondo. E questa struttura — si vuole mostrare in questo capitolo — non è limitata alla logica matematica: compare, in forme diverse ma strutturalmente identificabili, in ogni dominio in cui un sistema tenta di raggiungere la completezza rispetto a se stesso.

## II. La forma canonica dell'argomento diagonale

La formalizzazione più generale della costruzione diagonale è dovuta a Lawvere (1969), che la espresse nel linguaggio della teoria delle categorie in modo da renderla indipendente dagli specifici oggetti matematici su cui opera. Quella che segue è una riformulazione accessibile dell'argomento di Lawvere che ne preserva la struttura essenziale.

Siano  $A$  e  $Y$  due insiemi. Sia  $\varphi: A \rightarrow Y^A$  una funzione che mappa ogni elemento  $a \in A$  a una funzione  $\varphi_a: A \rightarrow Y$ . Si chiede: quando questa funzione può essere suriettiva — quando ogni funzione da  $A$  a  $Y$  può essere raggiunta come  $\varphi_a$  per qualche  $a$ ?

*Sia  $f: Y \rightarrow Y$  una qualsiasi funzione. Definiamo la funzione diagonale  $d: A \rightarrow Y$  come:*

$$d(a) = f(\varphi_a(a))$$

Questa è la "costruzione diagonale generalizzata" nel senso di Lawvere:  $d$  è definita applicando  $f$  al valore che  $\varphi_a$  assegna all'elemento  $a$  stesso — una forma di auto-applicazione mediata da  $f$ .

Teorema (Lawvere, 1969). Se  $\varphi: A \rightarrow Y^A$  è suriettiva, allora ogni funzione  $f: Y \rightarrow Y$  ha un punto fisso — esiste  $y^* \in Y$  tale che  $f(y^*) = y^*$ .

Dimostrazione. Poiché  $\varphi$  è suriettiva, la funzione diagonale  $d: a \rightarrow f(\varphi_a(a))$  è nell'immagine di  $\varphi$ . Quindi esiste  $a^* \in A$  tale che  $\varphi_{\{a^*\}} = d$  — ovvero  $\varphi_{\{a^*\}}(a) = f(\varphi_a(a))$  per ogni  $a \in A$ . In particolare, per  $a = a^*$ :

$$\varphi_{\{a^*\}}(a^*) = f(\varphi_{\{a^*\}}(a^*))$$

*Ponendo  $y^* = \varphi_{\{a^*\}}(a^*)$ , si ottiene  $y^* = f(y^*)$ .  $\square$*

Il teorema di Lawvere afferma: se  $A$  è abbastanza ricco da "rappresentare" tutte le funzioni da  $A$  a  $Y$  ( $\varphi$  è suriettiva), allora ogni funzione su  $Y$  ha un punto fisso. Il contropositivo è la forma operativamente utile: se qualche funzione  $f: Y \rightarrow Y$  non ha punto fisso, allora  $\varphi$  non può essere suriettiva —  $A$  non è abbastanza ricco da rappresentare tutte le funzioni da sé a  $Y$ .

Questo è il nucleo strutturale di tutti i teoremi diagonali:

Nel teorema di Cantor,  $Y = \{0,1\}$ ,  $A = \mathbb{N}$ ,  $\varphi$  è la supposta enumerazione dei reali, e  $f$  è la negazione ( $f(0) = 1$ ,  $f(1) = 0$ ). La negazione non ha punti fissi in  $\{0,1\}$  ( $0 \neq 1$  e  $1 \neq 0$ ), quindi  $\varphi$  non è suriettiva — l'enumerazione non copre tutti i reali.

Nella dimostrazione di Gödel,  $Y = \{\text{vero}, \text{falso}\}$ ,  $A$  = l'insieme delle sentenze di PA,  $\varphi$  è la funzione di soddisfacibilità ( $\varphi_a$  è la funzione che dice, per ogni sentenza  $b$ , se "a dice che b è vera"), e  $f$  è la negazione. La negazione non ha punto fisso nel dominio dei valori di verità (se  $\varphi_{\{a^*\}}(a^*) = \text{vero}$  allora  $f(\text{vero}) = \text{falso} \neq \text{vero}$ ), quindi il sistema non può rappresentare tutte le funzioni di verità — esistono sentenze vere che non sono catturate dalla funzione di dimostrabilità.

Nella dimostrazione di Turing,  $Y = \{\text{HALT}, \text{LOOP}\}$ ,  $A$  = l'insieme dei programmi,  $\varphi_p$  è la funzione del comportamento del programma  $p$  applicato agli input ( $\varphi_p(q) = \text{comportamento di } p \text{ su } q$ ), e  $f$  è la commutazione  $\text{HALT} \leftrightarrow \text{LOOP}$ . La commutazione non ha punto fisso (nessun programma che halta può risultare in loop e viceversa nel medesimo stato), quindi la supposizione che  $\varphi$  sia totalmente calcolabile è falsa — non esiste una macchina di Turing che calcoli il comportamento di ogni macchina di Turing su ogni input.

### III. Il teorema di Tarski come caso paradigmatico della Famiglia

#### IV

Il teorema di Tarski sull'indefinibilità della verità (1936) è il ponte tra la Famiglia I (impossibilità diagonali) e la Famiglia IV (impossibilità di chiusura semantica) identificate nel capitolo I. È simultaneamente dimostrato per diagonalizzazione e rivela un'impossibilità di auto-rappresentazione semantica.

Sia  $L$  un linguaggio formale che include l'aritmetica elementare. Sia  $\text{Tr}(x)$  un predicato ipotetico del linguaggio  $L$  tale che per ogni sentenza  $s$  di  $L$ :

$\text{Tr}(\ulcorner s \urcorner) \leftrightarrow s$  (la sentenza " $\text{Tr}(\ulcorner s \urcorner)$ " è vera se e solo se  $s$  è vera)

dove  $\ulcorner s \urcorner$  è il numero di Gödel di  $s$ . Questo predicato è il candidato per "verità nell'interno del linguaggio  $L$ " — se  $\text{Tr}(x)$  potesse essere definito in  $L$ ,  $L$  potrebbe esprimere la propria semantica.

**Teorema** (Tarski, 1936). Se  $L$  è consistente e sufficientemente ricco da contenere l'aritmetica elementare, allora non esiste alcun predicato  $\text{Tr}(x)$  di  $L$  che soddisfi lo schema  $T$  per tutte le sentenze di  $L$ .

**Dimostrazione.** Si usa il lemma della diagonalizzazione di Gödel: per qualsiasi predicato  $\varphi(x)$  di  $L$ , esiste una sentenza  $G$  tale che  $L \vdash G \leftrightarrow \varphi(\ulcorner G \urcorner)$ . Applicando il lemma al predicato  $\neg \text{Tr}(x)$ :

Esiste  $G$  tale che  $L \vdash G \leftrightarrow \neg \text{Tr}(\ulcorner G \urcorner)$

Ma per lo schema  $T$ :  $L \vdash \text{Tr}(\ulcorner G \urcorner) \leftrightarrow G$ . Sostituendo:  $L \vdash \text{Tr}(\ulcorner G \urcorner) \leftrightarrow \neg \text{Tr}(\ulcorner G \urcorner)$  — una contraddizione. Quindi il predicato  $\text{Tr}(x)$  non può essere definito in  $L$ .  $\square$

Il teorema di Tarski mostra che il confine tra un linguaggio e il suo metalinguaggio non è attraversabile dall'interno: il sistema  $L$  non può "guardare se stesso dall'esterno" con sufficiente distanza da vedere la propria verità. Qualsiasi tentativo produce la contraddizione della sentenza che dice di se stessa di non essere vera.

Questo è il nucleo della Famiglia IV: l'impossibilità di chiusura semantica non è una limitazione contingente — è la conseguenza necessaria del fatto che la semantica (il significato, la verità) non può essere interamente catturata dalla sintassi (la struttura formale) del medesimo sistema. Ogni volta che un sistema tenta questa cattura, produce o contraddizione o incompletezza.

#### IV. Il lemma della diagonalizzazione come generatore universale

Il lemma della diagonalizzazione — usato sia nella dimostrazione di Gödel sia in quella di Tarski — merita una formulazione autonoma perché è il meccanismo tecnico centrale di tutta la costruzione diagonale in logica. La sua struttura rivela perché l'auto-riferimento è irresistibile nei sistemi sufficientemente potenti.

Lemma della diagonalizzazione (Gödel, 1931). Sia  $T$  una teoria che contiene l'aritmetica elementare e sia  $\varphi(x)$  un predicato di  $T$  con una variabile libera. Allora esiste una sentenza  $G$  tale che:

$$T \vdash G \leftrightarrow \varphi(\ulcorner G \urcorner)$$

dove  $\ulcorner G \urcorner$  è il numero di Gödel di  $G$ .

L'interpretazione semantica è:  $G$  è una sentenza che "dice di sé" che ha la proprietà  $\varphi$ . Questo non è un'assunzione: è una conseguenza dimostrabile della struttura dell'aritmetica. La capacità di codificare i propri simboli come oggetti aritmetici (la "gödelizzazione") implica automaticamente la capacità di costruire sentenze auto-referenziali.

La dimostrazione del lemma usa la funzione di diagonalizzazione  $\text{diag}: \mathbb{N} \rightarrow \mathbb{N}$  definita come:

$\text{diag}(n)$  = il numero di Gödel della sentenza ottenuta sostituendo  $n$  nella formula con numero  $n$

*Questa funzione è calcolabile in  $T$ . Si definisce la sentenza  $D$  come:*

$$D = \varphi(\text{diag}(\ulcorner \varphi(\text{diag}(x)) \urcorner))$$

Allora  $G = D$  soddisfa il lemma. La costruzione è circolare in modo controllato:  $G$  dice di sé che ha la proprietà  $\varphi$ , e questo può essere espresso perché l'aritmetica ha abbastanza struttura da permettere la codifica di questa auto-referenza.

La conseguenza fondamentale è che in qualsiasi sistema abbastanza ricco da contenere l'aritmetica elementare, l'auto-riferimento non è una scelta — è una necessità. Il sistema non può essere progettato in modo da evitare le sentenze auto-referenziali: emergono automaticamente dalla struttura del sistema.

## V. La diagonalizzazione nelle impossibilità biologiche: le forme occulte

La tesi più originale di questo capitolo è che la costruzione diagonale non è limitata alla logica matematica e alla teoria della computazione: compare, in forme diverse ma strutturalmente riconoscibili, nelle impossibilità biologiche documentate nei due cicli precedenti. Il riconoscimento di queste "forme occulte" della diagonalizzazione richiede un'operazione di astrazione che questa sezione sviluppa sistematicamente.

La chiave dell'operazione è la seguente: ogni volta che un sistema tenta di rappresentare "se stesso" o "il suo ambiente" in modo completo, e ogni volta che quella rappresentazione viene valutata nel contesto che essa stessa descrive, si crea una struttura circolare analoga alla costruzione diagonale. Non è necessario che la circolarità sia esplicita o intenzionale: basta che emerga dalla struttura del problema.

La diagonalizzazione implicita nel binding problem. Il binding problem (Capitolo I del ciclo biologico) richiede di rappresentare, in uno spazio di dimensione  $n$ , la struttura di legame tra feature. L'impossibilità del binding esatto è dimostrabile attraverso un argomento diagonale nascosto.

Sia  $W \in \mathbb{R}^{n \times m}$  la matrice di embedding di  $m$  feature in  $n$  dimensioni ( $n < m$ ). L'operazione di binding produce un vettore  $v \in \mathbb{R}^n$  per ogni coppia di feature  $(f_i, f_j)$ :

$$v_{ij} = w_i \otimes w_j$$

dove  $\otimes$  è l'operazione di binding (ad esempio, la convoluzione circolare di Plate). Il recupero di  $f$  da  $v_{ij}$  dato  $w_i$  richiede un'operazione di de-binding:

$$\hat{w} = w_i \otimes^{-1} v_{ij}$$

Il problema sorge quando più oggetti sono rappresentati simultaneamente: sia  $V = \sum_i w_i \otimes w_i$  il vettore complessivo. Il recupero di  $w$  da  $V$  dato  $w_i$  produce:

$$\hat{w} = w_i \otimes^{-1} V = w_i \otimes^{-1} \left( \sum_k w_k \otimes w_k \right) = w + \sum_{\{(k,l) \neq (i,i)\}} w_i \otimes^{-1} (w_k \otimes w_l)$$

Il secondo termine è il "rumore" del recupero — il contributo dei legami degli altri oggetti. Per il recupero esatto, si richiede:

$$w_i \otimes^{-1} (w \otimes w) = 0 \text{ per } (k,l) \neq (i,i)$$

Questo richiede che  $w_i \otimes^{-1} w = 0$  per  $k \neq i$  — l'ortogonalità esatta di tutti i vettori di embedding. Ma l'ortogonalità esatta di  $m$  vettori in  $\mathbb{R}^n$  con  $m > n$  è impossibile: lo spazio non ha abbastanza dimensioni per contenere  $m$  vettori mutuamente ortogonali. La costruzione diagonale è implicita nella scelta del vettore di "rumore": si può costruire esplicitamente un pattern di binding tale che qualsiasi sistema di recovery commetta un errore, in analogia con la costruzione della sequenza diagonale di Cantor che non appartiene a nessuna enumerazione supposta completa.

La diagonalizzazione nel problema del frame. Il problema del frame richiede di specificare, per ogni azione  $a$  e ogni proprietà  $P$ , se  $a$  cambia  $P$ . La soluzione "banale" è elencare esplicitamente tutte le coppie  $(a, P)$  con effetto nullo — ma questo richiede un numero potenzialmente infinito di assiomi. Qualsiasi tentativo di ridurre questo numero attraverso uno schema generale ("tutto ciò che non è esplicitamente cambiato rimane invariato") genera un problema di auto-referenza.

La struttura diagonale emerge quando si considera la proprietà  $P_{\text{frame}} = \text{"essere la proprietà che questa azione non cambia"}$ . Questa proprietà è auto-referenziale: la sua applicazione a se stessa genera la circolarità:

$P_{\text{frame}} \text{ cambia} \leftrightarrow \text{"essere proprietà non cambiata"} \text{ è cambiata} \leftrightarrow P_{\text{frame}} \text{ non cambia}$

La circolarità non è un artefatto della formalizzazione — è la struttura stessa del problema del frame: qualsiasi schema generale per determinare ciò che rimane invariato deve applicarsi anche a se stesso, generando auto-riferimento.

La  $\Pi_2^P$ -completezza della Circumscription (che abbiamo citato nel capitolo I di questo ciclo come appartenente alla Famiglia III) è quindi anche una manifestazione della Famiglia I: l'intrattabilità computazionale del problema del frame è, in parte, una conseguenza dell'auto-riferimento diagonale che genera al suo interno.

La diagonalizzazione nel teorema No Free Lunch. Il teorema No Free Lunch (Capitolo II del ciclo biologico) afferma che nessun algoritmo di ottimizzazione supera la ricerca casuale quando si media su tutti i possibili paesaggi adattativi. L'argomento di Wolpert e Macready usa una struttura diagonale implicita.

Si supponga che l'algoritmo  $A$  superi la ricerca casuale sulla distribuzione uniforme  $F$  di tutti i paesaggi. Allora esiste un paecapitolo-"testimone"  $f^* \in F$  su cui  $A$  ottiene risultati migliori della media. Ma questo implica che  $A$  ha informazione a priori su  $f^*$  — che la sua struttura è "adattata" a  $f^*$  in qualche misura. Ora si costruisce il paecapitolo diagonale  $f_{\text{diag}}$  che "inverte"  $A$ : per ogni sequenza di mosse di  $A$  su qualsiasi paecapitolo,  $f_{\text{diag}}$  è il paecapitolo che produce esattamente il risultato opposto. Su  $f_{\text{diag}}$ ,  $A$  ha la peggior performance possibile. Poiché  $f_{\text{diag}}$  appartiene a  $F$ , e  $F$  è chiuso rispetto alle permutazioni (il punto chiave della dimostrazione di Wolpert-Macready), la performance media di  $A$  su  $F$  non può essere superiore a quella casuale.

La "costruzione inversa" del paecapitolo diagonale  $f_{\text{diag}}$  è strutturalmente identica alla costruzione cantoniana della sequenza diagonale: si costruisce un oggetto che "sfida" ogni elemento del supposto sistema completo, differendo da ciascun elemento nella dimensione rilevante.

## **VI. La costruzione diagonale nella biologia cognitiva: il paradosso dell'auto-osservazione**

L'analisi precedente ha identificato forme occulte della diagonalizzazione nelle impossibilità biologiche. Esiste però un caso ancora più diretto in cui la costruzione diagonale emerge nella biologia cognitiva: il problema dell'auto-osservazione — la capacità di un sistema cognitivo di osservare se stesso.

Qualsiasi sistema cognitivo  $S$  che tenti di costruire una rappresentazione completa di se stesso si trova in una situazione strutturalmente identica a quella di un sistema formale che tenta di contenere la propria funzione di verità. Formalmente: sia  $R(S)$  la rappresentazione interna di  $S$  a se stesso. Per essere "completa",  $R(S)$  deve contenere una rappresentazione di  $R(S)$  stessa — perché  $R(S)$  è parte di  $S$ , e  $S$  deve contenere tutto. Ma una rappresentazione di  $R(S)$  è una rappresentazione di una rappresentazione di  $S$  — e così via in regressione infinita.

La formalizzazione usa il teorema del punto fisso di Kleene, che è la versione computazionale del lemma della diagonalizzazione. Sia  $\varphi_e$  il programma con indice  $e$  (nella numerazione standard delle macchine di Turing). Per qualsiasi funzione computabile  $f: \mathbb{N} \rightarrow \mathbb{N}$ , esiste un indice  $e^*$  tale che:

$$\varphi_{-}\{f(e^*)\} = \varphi_{-}\{e^*\}$$

Il programma  $e^*$  è un punto fisso computazionale: applicare  $f$  al suo indice non cambia il comportamento del programma risultante. Applicato al problema dell'auto-osservazione: sia  $f$  la funzione che "aggiorna" la rappresentazione di  $S$  in risposta alla propria auto-osservazione. Il teorema del punto fisso garantisce che esiste uno stato di  $S$  in cui l'auto-osservazione non produce cambiamento — un stato di "auto-conoscenza stabile". Ma questo non è uno stato di conoscenza completa: è uno stato in cui la circolarità è neutralizzata dall'immobilità, non risolta.  $S$  "conosce se stesso" soltanto nel senso in cui una sentenza gödeliana  $G$  "sa" di essere indimostrabile — senza poter uscire dalla propria struttura formale per verificarlo dall'esterno.

Questa struttura si manifesta nel cervello come il problema dell'introspezione. Il sistema cognitivo che tenta di osservare i propri processi cognitivi in atto non può farlo completamente: l'osservazione stessa è un processo cognitivo che modifica i processi che si sta osservando (il "disturbo di osservazione" introspettivo), e qualsiasi descrizione di sé che il sistema produce è parziale — cattura ciò che il sistema può vedere di se stesso dallo stesso livello di descrizione. Il choice blindness documentato da Johansson et al. (discusso nel capitolo VI del ciclo biologico) è la manifestazione empirica di questo limite: il sistema produce spiegazioni post-hoc delle proprie scelte che non corrispondono ai processi che le hanno generate, proprio perché i processi che generano le scelte (Sistema 1) e i processi che producono le spiegazioni (Sistema 2) operano a livelli diversi della gerarchia, e il Sistema 2 non ha accesso diretto alla struttura del Sistema 1.

## **VII. Il teorema di Gödel come caso biologico: la mente non può dimostrare la propria consistenza**

Un'implicazione della costruzione diagonale che i due cicli precedenti non hanno sviluppato esplicitamente è la versione biologica del secondo teorema di incompletezza di Gödel: così come PA non può dimostrare la propria consistenza al proprio interno, il sistema cognitivo non può dimostrare la propria coerenza al proprio interno.

Il secondo teorema di Gödel, formalizzato da Hilbert-Bernays e poi da Löb (1955), afferma:

$T \vdash \text{Con}(T) \rightarrow \perp$  (se  $T$  è consistente,  $T$  non può dimostrare la propria consistenza)

dove  $\text{Con}(T)$  è la sentenza che esprime " $T$  è consistente" in termini gödeliani. Il risultato è che qualsiasi sistema abbastanza potente da includere l'aritmetica non può verificare la propria consistenza dall'interno — la verifica richiederebbe risorse che eccedono quelle del sistema stesso.

La traduzione biologica di questo risultato usa la struttura del framework del predictive coding (Capitolo VIII del ciclo biologico). Il modello generativo del cervello  $G$  è un sistema che rappresenta il mondo e genera previsioni. La "consistenza" del modello è la sua coerenza interna — il fatto che le previsioni generate in regioni diverse del modello non si contraddicano. Ma la verifica della consistenza del modello  $G$  richiederebbe un meta-modello  $G'$  che contenesse  $G$  e fosse in grado di verificarne la coerenza.  $G'$  dovrebbe a sua volta essere consistente, richiedendo  $G''$ , e così via.

Il cervello risolve praticamente questo problema non attraverso la dimostrazione della propria consistenza ma attraverso il meccanismo dell'energia libera variazionale: invece di verificare che il modello è consistente, minimizza l'energia libera  $F$  — che misura la sorpresa del modello rispetto ai dati — in modo tale che le inconsistenze producano alta energia libera e vengono progressivamente ridotte. Ma questa è una soluzione approssimata, non una garanzia: il sistema può raggiungere minimi locali di  $F$  che corrispondono a modelli internamente inconsistenti senza rilevare la propria inconsistenza. Le credenze deliranti (discusse nel capitolo VIII del ciclo biologico come aberrazioni della precisione) sono precisamente questo: stati del sistema in cui il modello è internamente inconsistente ma ha raggiunto un minimo locale di  $F$  che il sistema non può verificare come falso dall'interno.

## VIII. La forma biologica del lemma della diagonalizzazione:

### **l'evoluzione non può ottimizzarsi globalmente**

Il lemma della diagonalizzazione, nel suo significato biologico, afferma che qualsiasi sistema di ricerca sufficientemente espressivo genererà inevitabilmente strutture auto-referenziali che il sistema non può ottimizzare globalmente. In biologia, questo si manifesta come l'impossibilità dell'ottimizzazione evolutiva globale — una conseguenza che emerge dall'applicazione della struttura diagonale al paecapitolo adattativo.

Sia  $F: G \rightarrow \mathbb{R}$  la funzione di fitness (il "paecapitolo adattativo") che mappa ogni genotipo  $g \in G$  al suo valore di fitness. L'evoluzione per selezione naturale è un algoritmo di ascesa sul paecapitolo: partendo da una posizione  $g_0$ , si muove nelle direzioni di fitness crescente. L'impossibilità di ottimizzazione globale — il fatto che l'evoluzione rimane tipicamente intrappolata in massimi locali — è un caso speciale del teorema No Free Lunch (già discusso), ma ha una struttura diagonale aggiuntiva.

La struttura diagonale emerge considerando l'evoluzione come sistema che tenta di ottimizzare  $F$  valutando  $F$ . Ma  $F$  è definita, in parte, dall'ambiente biotico — dalla composizione dei conspecifici, dei predatori, delle prede, dei parassiti. E la composizione dell'ambiente biotico dipende dall'esito dell'evoluzione — il che rende  $F$  essa stessa una funzione dell'output del sistema di ottimizzazione. Il paecapitolo adattativo cambia in risposta alle mosse evolutive, e la funzione di fitness si trasforma ogni volta che l'evoluzione compie un passo.

Formalmente, sia  $F_t(g)$  la fitness del genotipo  $g$  al tempo  $t$ , con la dipendenza temporale data dalla dinamica coevolutiva:

$$dF_t/dt = \Phi(F_t, p_t)$$

dove  $p_t$  è la distribuzione dei genotipi nella popolazione al tempo  $t$ . Il sistema di ottimizzazione deve ottimizzare  $F_t$  al tempo  $t$ , ma  $F_t$  è essa stessa modificata dall'ottimizzazione. Questo genera una forma di auto-riferimento: il sistema ottimizza una funzione che è in parte definita dall'azione del sistema stesso.

La costruzione diagonale mostra che questo sistema non può convergere a un ottimo globale per qualsiasi funzione  $F$ . La prova è analoga a quella del teorema No Free Lunch: per qualsiasi strategia evolutiva  $A$ , si può costruire un ambiente

biottico "avversariale" che produce esattamente le pressioni selettive opposte a quelle che A è progettata per navigare — una versione biologica del paecapitolo diagonale.

## IX. La struttura categoriale: il teorema di Lawvere rivisitato

Il legame tra tutte le forme della diagonalizzazione — dal teorema di Cantor al ratchet di Müller, dall'incompletezza di Gödel al paradosso di Rogers — è catturato nella sua forma più generale dal teorema di Lawvere (1969) nella formulazione categoriale, che questa sezione sviluppa con la precisione tecnica necessaria.

In una categoria  $C$  con prodotti cartesiani, un oggetto  $A$  ha la proprietà del punto fisso se per ogni endomorfismo  $f: A \rightarrow A$  esiste un elemento  $x: 1 \rightarrow A$  tale che  $f \circ x = x$  (dove  $1$  è l'oggetto terminale). Un oggetto  $Y$  è un "classificatore di morfismi" per  $A$  se esiste una mappa di valutazione  $ev: Y^A \times A \rightarrow Y$  tale che per ogni morfismo  $g: B \times A \rightarrow Y$  esiste un unico morfismo  $\tilde{g}: B \rightarrow Y^A$  con  $ev \circ (\tilde{g} \times id_A) = g$ .

Teorema di Lawvere (forma categoriale). In una categoria cartesiana chiusa, se esiste un morfismo  $\varphi: A \rightarrow Y^A$  che è un epimorfismo (suriiettivo sulle componenti), allora  $Y$  ha la proprietà del punto fisso.

Il contropositivo fornisce la forma operativa: se  $Y$  non ha la proprietà del punto fisso (esiste  $f: Y \rightarrow Y$  senza punto fisso), allora nessun morfismo  $\varphi: A \rightarrow Y^A$  è un epimorfismo — l'oggetto  $A$  non può "rappresentare" completamente tutte le funzioni da  $A$  a  $Y$ .

Le istanze concrete di questo teorema nelle diverse famiglie di impossibilità sono:

In logica (Gödel):  $A = \mathbb{N}$  (i numeri naturali come codici delle sentenze),  $Y = \{\vdash, \nvdash\}$  (dimostrabile o non dimostrabile),  $\varphi =$  la funzione di dimostrabilità. L'endomorfismo  $f = \neg$  (negazione) non ha punto fisso in  $\{\vdash, \nvdash\}$ , quindi  $\varphi$  non è suriettiva — PA non cattura tutte le funzioni di verità.

In computabilità (Turing):  $A = \mathbb{N}$  (indici dei programmi),  $Y = \{\text{HALT}, \text{LOOP}\}$ ,  $\varphi_e =$  il comportamento del programma  $e$ . L'endomorfismo  $f = \text{switch}$  non ha punto fisso, quindi  $\varphi$  non è totalmente calcolabile.

Nella rappresentazione neurale (binding):  $A = \mathbb{R}^n$  (lo spazio delle attivazioni neurali),  $Y = \mathbb{R}^m$  con  $m > n$  (lo spazio delle feature),  $\varphi$  = la matrice di embedding  $W$ . L'endomorfismo  $f$  = riflessione non ha punto fisso fuori dall'origine, e per  $m > n$  l'embedding non è isometrico — qualche feature viene "compresa".

Nell'evoluzione (No Free Lunch):  $A$  = Algoritmi di ottimizzazione,  $Y$  = Prestazioni sull'insieme dei paesaggi,  $\varphi_A$  = la funzione di performance di  $A$ . L'endomorfismo  $f$  = inversione della prestazione non ha punto fisso, quindi nessun algoritmo  $A$  può dominare uniformemente su tutti i paesaggi.

La struttura categoriale unifica queste istanze diverse sotto una proprietà formale comune: in tutti i casi, l'impossibilità emerge dall'esistenza di un endomorfismo senza punto fisso nello spazio dei valori  $Y$ , combinata con il tentativo di costruire una rappresentazione  $\varphi$  completa dallo spazio  $A$  allo spazio delle funzioni  $Y^A$ .

## **X. Il significato della struttura diagonale: cosa ci dice dell'universo**

La domanda finale che questo capitolo pone — e alla quale le successive risponderanno gradualmente — è: perché la costruzione diagonale è così universalmente applicabile? Non è sufficiente osservare che funziona in molti contesti: occorre capire perché funziona, qual è la struttura profonda del mondo che la rende efficace.

La risposta provvisoria che emerge dall'analisi sviluppata in questo capitolo è la seguente: la costruzione diagonale funziona ogni volta che un sistema tenta di raggiungere la completezza rispetto a se stesso — ogni volta che il sistema deve essere sia l'oggetto descritto sia il descrittore, sia il soggetto dell'ottimizzazione sia il contesto dell'ottimizzazione, sia la struttura rappresentata sia il mezzo della rappresentazione.

Questo tentativo di completezza auto-referenziale non è una scelta — è la struttura stessa di qualsiasi sistema sufficientemente potente. Un sistema abbastanza ricco da descrivere il proprio ambiente è necessariamente abbastanza ricco da tentare di descrivere se stesso. E questa auto-descrizione, come il teorema di Lawvere formalizza in modo definitivo, non può essere completa.

La costruzione diagonale è quindi la traccia matematica di un principio più profondo che potremmo chiamare principio di incompiutezza

dell'auto-riferimento: ogni sistema che tenta di contenere una rappresentazione di se stesso nell'atto del contenere trova un bordo — un punto oltre il quale la rappresentazione non può essere completata senza generare contraddizione o incompletezza.

Questo principio non è una limitazione accidentale dei sistemi che lo esibiscono. È la condizione sotto cui quei sistemi sono interessanti — la condizione sotto cui possono produrre novità, creatività, sorpresa. Un sistema completamente chiuso su se stesso — senza questo bordo di incompletezza — sarebbe un cristallo, non una mente. La costruzione diagonale è ciò che mantiene aperto il sistema. Dove inizia l'impossibilità, inizia anche la possibilità di ciò che non è ancora stato pensato.

---

## Capitolo III

### Il confine sintassi-semantica come locus generativo

#### Perché tutte le impossibilità strutturali convergono verso un unico bordo

##### Mapa epistemica — Capitolo III

| Status          | Cod. | Claim                                                                                                                                                                 |
|-----------------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | La distinzione formale sintassi/semantica (Tarski 1933) e il teorema di Löwenheim-Skolem (1920) sono risultati formalmente dimostrati.                                |
| CONGETTURA      | C    | Versione forte: il confine sintassi-semantica è la causa generativa di tutte le impossibilità strutturali. Congettura fondata, non teorema.                           |
| INTERPRETAZIONE | I    | L'applicazione del confine alle impossibilità biologiche (paradosso di Rogers, divario esplicativo, wanting/liking): proposte interpretative strutturalmente fondate. |

#### I. La promessa da mantenere

nel capitolo I di questo ciclo si è avanzata una tesi che richiede ora di essere sostenuta con la massima precisione: tutte le impossibilità strutturali catalogate nei due cicli precedenti — indipendentemente dalla loro famiglia di appartenenza e dal dominio in cui emergono — sono manifestazioni del confine tra sintassi e semantica. Questo confine non è una metafora linguistica applicata a fenomeni eterogenei: è una struttura formale precisa la cui ubiquità è dimostrabile. Il capitolo II ha già mostrato come le impossibilità diagonali siano casi di fallimento dell'auto-chiusura semantica; questo capitolo estende l'analisi alle famiglie rimanenti e, soprattutto, argomenta perché il confine sintassi-semantica occupa la posizione di locus generativo — non semplicemente uno dei molti modi in cui le impossibilità si manifestano, ma la struttura che le genera tutte.

La tesi ha una versione debole e una versione forte. La versione debole afferma che il confine sintassi-semantica è presente in ogni impossibilità strutturale come componente necessaria della sua struttura. La versione forte afferma che il confine sintassi-semantica è la causa generativa delle impossibilità — che ogni impossibilità

strutturale può essere derivata come conseguenza dell'esistenza di questo confine. Questo testo sostiene la versione forte, con la consapevolezza che la piena dimostrazione richiederebbe un apparato formale che in parte appartiene ancora al futuro della ricerca.

## II. La distinzione sintassi-semantica: chiarimento preliminare

La distinzione tra sintassi e semantica — nata nella logica formale degli anni Trenta del Novecento attraverso il lavoro di Tarski, Carnap e Gödel — ha subito nel corso dei decenni una proliferazione di usi che rischia di svuotarla di contenuto preciso. Prima di usarla come strumento analitico, è necessario ridefinirla con la nitidezza che questo ciclo richiede.

Nel senso tecnico della logica formale, la sintassi di un linguaggio  $L$  è l'insieme delle sue regole di formazione e trasformazione: quali sequenze di simboli sono formule ben formate, quali inferenze sono derivazioni legittime, quali strutture sono prove valide. La sintassi opera esclusivamente sulle forme dei simboli — non sui loro significati. Una regola sintattica come "da  $A$  e  $A \rightarrow B$  si può derivare  $B$ " (modus ponens) è applicabile indipendentemente da cosa significhino  $A$  e  $B$ .

La semantica di  $L$  è l'interpretazione — la funzione che assegna significato alle espressioni di  $L$  in un dominio di riferimento. La semantica specifica cosa rendono vera una sentenza, quali strutture soddisfano una formula, quale valore ha un termine nel dominio. La semantica dipende da una corrispondenza tra i simboli del linguaggio e le entità del dominio che è esterna alla struttura sintattica.

La distinzione di Tarski ha una precisione tecnica che vale la pena recuperare. Per un linguaggio  $L$ , l'enunciato "la neve è bianca" è una formula sintattica — una stringa di simboli con certa struttura grammaticale. Che questa formula sia vera è un fatto semantico che non è contenuto nella struttura della formula ma nella corrispondenza tra la formula e il mondo: il predicato "è bianca" è soddisfatto dall'oggetto a cui "la neve" si riferisce. Questa corrispondenza — la relazione di soddisfazione — deve essere definita in un metalinguaggio che "parla di"  $L$  dall'esterno.

Il confine che questo capitolo studia è la linea che separa ciò che può essere detto nella sintassi da ciò che può essere detto nella semantica — o, più precisamente, la

linea oltre la quale la sintassi non può rappresentare la semantica senza generare contraddizione o incompletezza. Questo confine è il luogo in cui tutte le impossibilità strutturali si incontrano.

### III. La gerarchia di Tarski come topografia del confine

Alfred Tarski (1933, pubblicato in tedesco; 1956, traduzione inglese) costruì la soluzione al problema dell'auto-referenza semantica attraverso la gerarchia dei metalinguaggi. Il linguaggio oggetto  $L_0$  non può definire il proprio predicato di verità; ma un metalinguaggio  $L_1$  che "parla di"  $L_0$  può farlo.  $L_1$  a sua volta non può definire il proprio predicato di verità, ma un meta-metalinguaggio  $L_2$  può definire la verità di  $L_1$ . La gerarchia è:

$$L_0 \subset L_1 \subset L_2 \subset \dots$$

con  $L$  in grado di definire il predicato di verità per  $L_{-1}$ . Ogni livello della gerarchia cattura più semantica del livello precedente ma genera nuove verità semantiche che richiedono il livello successivo per essere catturate.

La struttura formale della gerarchia tarskiana è quella di una catena strettamente crescente di sistemi logici, dove ogni sistema estende il precedente con la capacità di esprimere la verità del precedente. Questa catena non ha un limite superiore finito: non esiste un livello  $L_\omega$  che catturi la semantica di tutti i livelli inferiori simultaneamente senza generare auto-riferimento.

Il rilevante per questo ciclo è che la gerarchia di Tarski non è soltanto una soluzione tecnica al paradosso del mentitore: è una topografia del confine sintassi-semantica. Ogni livello della gerarchia è un punto sul confine — un luogo in cui la sintassi del sistema corrente incontra il limite di ciò che può catturare della propria semantica. Il confine non è un muro singolo ma una successione infinita di soglie, ciascuna delle quali rivela la struttura di impossibilità del livello immediatamente precedente.

La topografia può essere formalizzata attraverso la nozione di "potenza espressiva" di un sistema. Sia  $\text{Exp}(L)$  la classe delle verità esprimibili in  $L$ . Allora:

$$\text{Exp}(L) \subsetneq \text{Exp}(L_{+1})$$

per ogni  $n$ , con la differenza  $\text{Exp}(L_{+1}) \setminus \text{Exp}(L)$  contenente esattamente le verità semantiche su  $L$  che  $L$  non può esprimere. Il confine sintassi-semantica al livello  $n$  è la frontiera tra  $\text{Exp}(L)$  e  $\text{Exp}(L_{+1})$  — l'insieme delle verità che richiedono il salto al livello superiore.

#### IV. La semantica come invarianza: una formalizzazione alternativa

La distinzione sintassi-semantica può essere formalizzata in modo alternativo — più ricco di implicazioni per le impossibilità biologiche — attraverso la nozione di invarianza. La sintassi di un sistema è ciò che rimane invariante rispetto alle trasformazioni puramente formali (le permutazioni dei simboli che preservano la struttura grammaticale ma non il contenuto). La semantica è ciò che è invariante rispetto alle trasformazioni del dominio di interpretazione (le mappe tra modelli che preservano la struttura di soddisfazione).

Questa formalizzazione usa il linguaggio della teoria dei modelli. Sia  $T$  una teoria (un insieme di assiomi) e sia  $\text{Mod}(T)$  la classe dei modelli di  $T$  (le strutture che soddisfano  $T$ ). Due modelli  $M$  e  $M'$  sono elementarmente equivalenti, scritto  $M \equiv M'$ , se soddisfano le stesse sentenze di  $T$ . Le proprietà sintattiche di  $T$  — quelle determinate dalla struttura formale degli assiomi — sono invarianti per equivalenza elementare: se  $M \equiv M'$  e  $T \vdash \varphi$ , allora sia  $M$  sia  $M'$  soddisfano  $\varphi$ . Le proprietà semantiche — quelle che dipendono dalla struttura specifica del modello, non dalla sua equivalenza con altri — non sono necessariamente invarianti per equivalenza elementare.

Il teorema di Löwenheim-Skolem (1920) è il risultato più eloquente di questa asimmetria. Qualsiasi teoria  $T$  con un modello infinito ha modelli di ogni cardinalità infinita — inclusa la cardinalità numerabile  $\aleph_0$ . In particolare, la teoria degli insiemi ZFC (che include assiomi che garantiscono l'esistenza di insiemi non numerabili come  $\mathbb{R}$ ) ha un modello numerabile — il modello di Skolem. Questo è il "paradosso di Skolem": la teoria che afferma l'esistenza di insiemi non numerabili ha un modello numerabile. Come è possibile?

La risoluzione è precisamente la distinzione sintassi-semantica: la non-numerabilità di  $\mathbb{R}$  è una proprietà semantica (dipende dalla struttura effettiva dei modelli), non sintattica (non è determinata dagli assiomi di ZFC). Nel modello di Skolem, esiste una biiezione tra  $\mathbb{R}$  e  $\mathbb{N}$  — ma quella biiezione non è nel modello (non è un

elemento del modello), e quindi la teoria non può "vederla" dall'interno. La teoria non può catturare la propria semantica: ciò che i suoi assiomi affermano sulla cardinalità dipende dal modello in cui si interpretano quegli assiomi.

Questa è la struttura di ogni impossibilità di chiusura semantica: non è che la semantica manchi — è che la sintassi non può raggiungerla. La sintassi parla della struttura; la semantica parla del significato di quella struttura in un dominio; e il dominio è sempre, in una certa misura, esterno al sistema che lo descrive.

## **V. Il confine sintassi-semantica nelle impossibilità biologiche: una fenomenologia sistematica**

La tesi centrale di questo capitolo richiede di identificare, in ogni impossibilità biologica catalogata nei cicli precedenti, la struttura specifica del confine sintassi-semantica che la genera. Questa sezione conduce questa identificazione in modo sistematico, organizzandola per famiglia.

Impossibilità della Famiglia I nelle forme biologiche. Il paradosso di Rogers (la trasmissione culturale pura non aumenta la fitness media della popolazione) è un caso di confine sintassi-semantica preciso. La sintassi dell'apprendimento sociale è la struttura osservabile del comportamento — la sequenza di mosse, di gesti, di azioni che il modello esegue. La semantica è l'adattamento all'ambiente specifico che quelle azioni implementano — il fatto che quella sequenza di azioni funzioni nel contesto ecologico in cui è stata prodotta. L'apprendimento sociale copia la sintassi (il comportamento osservabile) ma non necessariamente la semantica (la ragione per cui quel comportamento funziona in quell'ambiente). Quando l'ambiente cambia — come avviene nel modello di Rogers con tasso di cambiamento  $u > 0$  — la semantica diventa obsoleta mentre la sintassi rimane invariata: il comportamento copiato non è più adattivo. Questo è precisamente il paradosso di Rogers: la copia sintattica senza aggiornamento semantico non produce vantaggio adattivo medio.

Formalmente: sia  $B$  il comportamento e sia  $E$  l'ambiente. La fitness è  $F(B, E)$  — una funzione che dipende sia dal comportamento sia dall'ambiente (la semantica del comportamento rispetto all'ambiente). L'apprendimento sociale copia  $B$  ma non  $F(\cdot, E)$ , il che significa che copia la sintassi (la struttura del comportamento)

ma non la semantica (il valore del comportamento nell'ambiente corrente). Quando E cambia in E', il comportamento copiato B ha fitness  $F(B, E')$  che può essere molto diverso da  $F(B, E)$ . Il gap semantico è  $F(B, E) - F(B, E')$  — e questo gap è precisamente la perdita che il paradosso di Rogers quantifica.

Impossibilità della Famiglia II nelle forme biologiche. Il binding problem è già stato analizzato nel capitolo II come caso di costruzione diagonale implicita. Ma ha anche una struttura di confine sintassi-semantica che vale la pena rendere esplicita. La sintassi della rappresentazione neurale è la struttura vettoriale degli spazi di attivazione — la geometria degli spazi di embedding, le distanze e le direzioni. La semantica è il significato degli schemi di attivazione — il fatto che un certo pattern di attivazione corrisponde a "cerchio rosso a sinistra" invece che a "quadrato blu a destra". Il binding problem è precisamente l'impossibilità di conservare la struttura semantica (i legami tra feature distinte) nel passaggio alla struttura sintattica (la rappresentazione vettoriale in uno spazio di dimensione fissa). Il confine è il punto in cui la ricchezza semantica (tutti i possibili legami tra feature) non può essere completamente rappresentata nella sintassi (lo spazio vettoriale finito).

Impossibilità della Famiglia III nelle forme biologiche. La  $\Pi_2^P$ -completezza del problema del frame appartiene alla Famiglia III come impossibilità di complessità, ma la sua struttura più profonda è quella di un confine sintassi-semantica. La sintassi del frame problem è la struttura formale dei predicati sulle proprietà del mondo: quali proprietà sono soddisfatte in quali stati, quali azioni trasformano quali stati. La semantica è la rilevanza — quale delle infinite proprietà del mondo conta in questo contesto per questo agente con questi scopi. La rilevanza non è definibile sintatticamente: è una funzione del significato delle proprietà rispetto agli scopi dell'agente, non della struttura formale delle proprietà stesse. Qualsiasi tentativo di catturare la rilevanza in termini sintattici (come fa la Circumscription) genera la  $\Pi_2^P$ -completezza: il sistema deve verificare condizioni su tutte le possibili interpretazioni semantiche del contesto.

Il divario esplicativo come confine puro. Il divario esplicativo di Chalmers (Capitolo IV del ciclo biologico) è il caso più diretto e più trasparente di impossibilità di chiusura semantica. La fisica è, esplicitamente nel programma di Chalmers, una descrizione sintattica del mondo: descrive strutture relazionali,

proprietà matematiche, leggi di trasformazione — la forma del mondo senza il suo contenuto qualitativo. I qualia sono, esplicitamente, proprietà semantiche: si riferiscono all'esperienza soggettiva, al significato che certi stati fisici hanno "dall'interno" di un sistema che li attraversa. Il divario esplicativo è la distanza tra la sintassi fisica e la semantica fenomenica — e il teorema implicito che quella distanza non è colmabile dalla fisica è esattamente l'impossibilità tarskiana: la fisica (il linguaggio  $L_0$ ) non può definire il predicato di verità per l'esperienza soggettiva (la semantica di  $L_0$ ) dall'interno.

## VI. La struttura formale del confine: la relazione di soddisfazione come oggetto irraggiungibile

La relazione di soddisfazione — la relazione semantica centrale della logica formale — è l'oggetto che il confine sintassi-semantica rende irraggiungibile dall'interno. Questa sezione ne sviluppa la struttura formale in modo da mostrare perché è irraggiungibile e in che senso è il generatore di tutte le impossibilità.

Sia  $L$  un linguaggio formale del primo ordine con formule  $\varphi$ , e sia  $M = (D, I)$  un modello ( $D$  è il dominio degli oggetti,  $I$  è la funzione di interpretazione dei simboli non logici). La relazione di soddisfazione  $\models$  è definita recursivamente:

$$\begin{aligned} M \models P(t_1, \dots, t_n) &\text{ sse } I(P)(I(t_1), \dots, I(t_n)) \text{ è vero in } D \\ M \models \neg \varphi &\text{ sse non } M \models \varphi \\ M \models \varphi \wedge \psi &\text{ sse } M \models \varphi \text{ e } M \models \psi \\ M \models \forall x \varphi &\text{ sse per ogni } d \in D, M[x/d] \models \varphi \end{aligned}$$

Questa definizione è semantica — è definita in termini del dominio  $D$  e della funzione di interpretazione  $I$ , che sono strutture esterne al linguaggio  $L$ .  $L$  non può parlare di  $D$  e  $I$  dall'interno: può soltanto usare simboli che vengono interpretati in  $D$  attraverso  $I$ .

Il tentativo di catturare  $\models$  all'interno di  $L$  genera il paradosso del mentitore:

Sia  $\text{Sat}(x, y)$  un predicato di  $L$  che si suppone esprima "la formula con numero di Gödel  $x$  è soddisfatta dall'oggetto con numero di Gödel  $y$ ". Il lemma della diagonalizzazione produce una sentenza  $G$  tale che:

$$L \vdash G \leftrightarrow \neg \text{Sat}(\ulcorner G \urcorner, \ulcorner M \urcorner)$$

$G$  afferma di non essere soddisfatta nel modello  $M$ . Se  $G$  è vera nel modello  $M$ , allora  $M \models G$ , ma allora per lo schema di soddisfazione  $M \models \neg \text{Sat}(\ulcorner G \urcorner, \ulcorner M \urcorner)$ , il che

contraddice  $M \models G$ . Se  $G$  è falsa, il ragionamento simmetrico produce l'altra contraddizione. Il predicato  $\text{Sat}(x, y)$  non può essere definito in  $L$ .

Il teorema di Tarski è la versione di questo argomento per il predicato di verità: il predicato  $\text{Tr}(x) = \text{Sat}(x, \ulcorner M \urcorner)$  non è definibile in  $L$ . La relazione di soddisfazione è l'oggetto irraggiungibile — il bordo del confine sintassi-semantica.

Ma ora si può vedere qualcosa di più preciso: il confine sintassi-semantica non è un confine singolo ma è la coppia  $(L, \models)$ . La sintassi è  $L$  — la struttura formale del linguaggio. La semantica è  $\models$  — la relazione che connette le strutture formali ai significati. Il confine è la relazione stessa: né puramente sintattica (non è definibile in  $L$ ) né esterna al sistema (è la struttura che connette  $L$  al suo dominio). Il confine è la relazione di soddisfazione — e questa relazione è irraggiungibile dall'interno di  $L$  per il teorema di Tarski.

## **VII. Il confine come struttura produttiva: i limiti che generano conoscenza**

La versione debole della tesi — il confine sintassi-semantica è presente in ogni impossibilità strutturale — è stata sostenuta con esempi sistematici. La versione forte — il confine è la causa generativa delle impossibilità — richiede un argomento più sottile.

L'argomento per la versione forte procede in tre passi.

Il primo passo è mostrare che la distinzione sintassi-semantica è necessariamente presente in qualsiasi sistema sufficientemente espressivo. Un sistema  $S$  è sufficientemente espressivo se può descrivere elementi non banali del proprio ambiente. Ma descrivere l'ambiente richiede che i simboli di  $S$  abbiano un referente — un significato — nel dominio  $D$  del sistema. E la relazione tra i simboli (sintassi) e i referenti (semantica) è necessariamente esterna alla struttura formale di  $S$ : non può essere interamente specificata dai simboli stessi, perché qualsiasi specificazione sintattica della semantica sarebbe essa stessa una struttura formale che richiede a sua volta un'interpretazione. La distinzione sintassi-semantica è quindi una proprietà inevitabile di qualsiasi sistema abbastanza ricco da descrivere qualcosa.

Il secondo passo è mostrare che la presenza del confine genera impossibilità. Il meccanismo è quello del teorema di Lawvere: un sistema  $S$  con sintassi e semantica distinte non può costruire una mappa  $\varphi: S \rightarrow \text{Sem}(S)^S$  che sia suriettiva (dove  $\text{Sem}(S)$  è lo spazio dei valori semantici). Se potesse, la funzione di negazione semantica ( $f: \text{Sem}(S) \rightarrow \text{Sem}(S)$  che inverte il valore di verità) non avrebbe punto fisso, e la surgettività di  $\varphi$  implicherebbe l'esistenza di un punto fisso per  $f$  — contraddizione. Quindi  $\varphi$  non è suriettiva — il sistema non può rappresentare completamente tutte le relazioni semantiche.

Il terzo passo è mostrare che ogni impossibilità strutturale è un'istanza di questa relazione tra confine e incompletezza. Ogni impossibilità strutturale si presenta come il fallimento di un certo tipo di completezza (della separabilità lineare, della decisione dell'arresto, del binding esatto, della trasmissione culturale fedele) — e ogni tipo di incompletezza è la traccia di un confine sintassi-semantica specifico (la dimensione dello spazio di rappresentazione, la struttura del modello computazionale, la capacità dello spazio di embedding, la fedeltà della copia comportamentale).

## VIII. L'asimmetria informazionale come misura del confine

Se il confine sintassi-semantica è la struttura che genera le impossibilità, deve esistere una misura quantitativa della distanza tra sintassi e semantica che corrisponda alla severità dell'impossibilità. Questa sezione propone una tale misura attraverso il concetto di gap informazionale.

Sia  $S$  un sistema con struttura sintattica  $\Sigma$  e struttura semantica  $M$  (un modello). Il gap informazionale è la quantità di informazione contenuta in  $M$  che non è catturata da  $\Sigma$ :

$$\text{Gap}(\Sigma, M) = H(M) - I(M; \Sigma)$$

dove  $H(M)$  è l'entropia di Shannon del modello (la quantità di informazione necessaria per specificare  $M$ ) e  $I(M; \Sigma)$  è l'informazione mutua tra il modello e la struttura sintattica (la quantità di informazione su  $M$  che è inferibile da  $\Sigma$ ). Il gap è sempre non-negativo (per il teorema della subadditività dell'entropia) e si annulla soltanto quando la sintassi cattura completamente la semantica — che per il teorema di Tarski è impossibile per sistemi sufficientemente ricchi.

Per le impossibilità della Famiglia II, il gap informativo ha una forma particolarmente esplicita. Nella rappresentazione neurale di  $m$  feature in  $n < m$  dimensioni, la sintassi è lo spazio delle attivazioni  $\mathbb{R}^n$  e la semantica è lo spazio delle feature  $\mathbb{R}^m$ . Il gap informativo è:

$$\text{Gap}(\mathbb{R}^n, \mathbb{R}^m) = H(\mathbb{R}^m) - I(\mathbb{R}^m; \mathbb{R}^n) \geq (m - n) \log(2\pi e \sigma^2)$$

per feature con varianza  $\sigma^2$ , dove il bound inferiore è l'informazione nelle  $(m - n)$  dimensioni dello spazio delle feature che non sono catturate dallo spazio delle attivazioni. Ogni dimensione "mancante" corrisponde a un grado di libertà semantico che la sintassi non può rappresentare — un contributo al gap che produce interferenze nel binding.

Per le impossibilità della Famiglia III, il gap informativo è legato alla profondità nella gerarchia polinomiale. Sia  $\Pi_2^P$  il livello della Circumscription. Il gap informativo tra la sintassi del sistema logico (i predicati di  $L_0$ ) e la semantica del ragionamento non-monotono (i modelli di Circumscription) è almeno del secondo livello della gerarchia: capire cosa è "minimalmente cambiato" da un'azione richiede quantificare su tutti i modelli possibili — un livello semantico che eccede la sintassi di primo livello del sistema.

Per le impossibilità della Famiglia IV — la chiusura semantica — il gap informativo è infinito nel senso della gerarchia tarskiana: nessuna quantità finita di estensione sintattica cattura completamente la semantica, perché ogni estensione genera nuove verità semantiche che richiedono un'ulteriore estensione.

## IX. Il confine sintassi-semantica come struttura universale:

### l'argomento della necessità

La tesi finale di questo capitolo — la versione forte — richiede di mostrare che il confine sintassi-semantica è necessariamente presente in qualsiasi sistema che esibisce impossibilità strutturali. L'argomento procede per contropositivo: si mostra che un sistema privo del confine sintassi-semantica non può essere abbastanza ricco da porre le domande a cui le impossibilità strutturali rispondono negativamente.

Un sistema  $S$  è "piatto" (privo del confine sintassi-semantica) se la sua struttura sintattica coincide completamente con la sua struttura semantica — se ogni

relazione semantica in  $S$  è esprimibile con i simboli di  $S$  senza residuo. Esempi di sistemi piatti: l'algebra booleana su  $\{0,1\}$ , la logica proposizionale su un insieme finito fisso di proposizioni, i sistemi di regole finite senza quantificatori. In questi sistemi, ogni verità è decidibile, ogni inferenza è calcolabile in tempo finito, ogni rappresentazione è esatta.

Ma questi sistemi piatti non sono abbastanza ricchi da descrivere l'aritmetica degli interi, il comportamento delle macchine di Turing, la struttura degli spazi di Hilbert di dimensione infinita, la dinamica evolutiva delle popolazioni con geni arbitrari. La ricchezza necessaria per descrivere questi oggetti richiede sistemi quantificati (per descrivere "per ogni intero  $n...$ "), sistemi di auto-riferimento (per descrivere il comportamento di un programma su se stesso), sistemi di dimensione infinita (per descrivere gli spazi funzionali). E sistemi con queste proprietà non possono essere piatti: il teorema di Tarski garantisce che qualsiasi sistema con auto-riferimento e quantificazione sufficiente avrà un confine sintassi-semantica.

L'argomento formale è questo: sia  $S$  un sistema abbastanza ricco da porre la domanda " $S$  è completo rispetto alla propria semantica?" — ovvero, da esprimere la propria semantica come oggetto matematico. Allora  $S$  contiene, come sottosistema, abbastanza aritmetica da applicare il lemma della diagonalizzazione, e quindi il teorema di Tarski si applica:  $S$  non può definire il proprio predicato di verità. Il confine sintassi-semantica è presente. Se invece  $S$  non è abbastanza ricco da porre questa domanda, non è abbastanza ricco da essere il tipo di sistema a cui si applicano le impossibilità strutturali che stiamo studiando.

Il confine sintassi-semantica è quindi sia necessario sia sufficiente per la presenza di impossibilità strutturali: necessario perché qualsiasi sistema abbastanza ricco lo possiede, sufficiente perché la sua presenza genera le impossibilità attraverso il meccanismo di Lawvere. Il confine sintassi-semantica è il locus generativo — la struttura da cui tutte le impossibilità emanano.

## **X. Le conseguenze per la comprensione del cervello biologico**

La tesi del confine sintassi-semantica come locus generativo ha implicazioni dirette per la comprensione del cervello biologico che vale la pena sviluppare esplicitamente, anche se la loro elaborazione piena appartiene ai capitoli successive.

Il cervello biologico opera con strutture sia sintattiche sia semantiche. Le strutture sintattiche sono i pattern di attivazione neurale — la geometria dello spazio delle attivazioni, i ritmi oscillatori, la topologia del connettoma. Le strutture semantiche sono i significati di quei pattern — il fatto che un certo pattern di attivazione in IT corrisponde alla percezione di un volto, che un certo segnale dopaminergico corrisponde alla sorpresa di una ricompensa inattesa, che un certo pattern di attività prefrontale corrisponde alla pianificazione di un'azione.

Il confine sintassi-semantica nel cervello è il punto in cui la struttura neurale (sintattica) si connette al significato (semantico). Questo confine non è fisso: cambia con l'apprendimento, con il contesto, con lo sviluppo. Ma il teorema di Tarski implica che, per qualsiasi stato del cervello, esiste sempre un gap tra la struttura neurale e il suo significato — un'area di "semantica irraggiunta" che la struttura neurale corrente non cattura.

Il principio di energia libera di Friston (Capitolo VIII del ciclo biologico) è un tentativo di attraversare questo confine dall'interno: il cervello minimizza l'energia libera variazionale  $F$ , che misura la discrepanza tra il modello generativo (la sintassi delle previsioni) e i dati sensoriali (la semantica dell'esperienza). Ma  $F$  è misurabile soltanto dal livello sintattico — dalla struttura del modello generativo — e la minimizzazione di  $F$  non equivale alla minimizzazione del gap semantico. Il gap tra la struttura del modello ( $F$  bassa) e il significato del modello (la verità delle sue previsioni nel mondo) rimane irriducibile.

Questa è la versione biologica del teorema di Tarski: il cervello non può verificare la propria semantica dall'interno — può soltanto ottimizzare la propria sintassi nella direzione di una migliore corrispondenza con il segnale sensoriale. Il significato reale delle previsioni del cervello — se le sue rappresentazioni del mondo sono vere, se la sua mappa corrisponde al territorio — è sempre parzialmente opaco al cervello stesso. Questa opacità non è un difetto correggibile: è la conseguenza strutturale del confine sintassi-semantica, e come tale è condivisa da tutti i sistemi abbastanza ricchi da avere un modello del mondo.

## Capitolo IV

### Le impossibilità dimensionali e la geometria della compressione

#### La struttura quantitativa dei limiti di rappresentazione e il loro fondamento nella topologia degli spazi metrici

##### Mappe epistemiche — Capitolo IV

| Status          | Cod. | Claim                                                                                                                                                                         |
|-----------------|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | Teorema JL (1984) e il suo lower bound (Alon 2003; Jayram-Woodruff 2013): formalmente dimostrati.                                                                             |
| DIMOSTRATO      | D    | Formula di Cover per la separabilità lineare e connessione con la dimensione VC: dimostrata.                                                                                  |
| INTERPRETAZIONE | I    | Il binding problem neurobiologico come istanza biologica del converse di JL; la codifica sparsa come soluzione ottimale sulla frontiera: proposta interpretativa ben fondata. |

#### I. Una famiglia trascurata

Tra le quattro famiglie di impossibilità strutturali identificate nel capitolo I di questo ciclo, la Famiglia II — le impossibilità dimensionali — è quella che ha ricevuto meno attenzione filosofica sistematica, nonostante sia la più direttamente rilevante per la biologia della cognizione e la più ricca di struttura quantitativa. Le impossibilità diagonali hanno una tradizione intellettuale potente da Cantor a Gödel; le impossibilità di complessità hanno una letteratura tecnica sterminata nella teoria della complessità computazionale; le impossibilità di chiusura semantica hanno il peso della filosofia analitica del Novecento. Le impossibilità dimensionali sono state studiate intensamente in matematica applicata e in machine learning, ma raramente il loro significato strutturale è stato portato al centro della riflessione epistemologica.

Eppure le impossibilità dimensionali hanno una proprietà che le altre famiglie non condividono: sono continue e quantitative. Non dicono semplicemente "impossibile" o "possibile" — dicono "possibile con errore  $\epsilon$  in funzione del

rapporto  $m/n$  tra la dimensione del problema e la dimensione della rappresentazione". Questa struttura quantitativa le rende calcolabili, predittive, e ingegneristicamente utili in un senso che le impossibilità diagonali e di chiusura semantica non possono essere. E questa stessa struttura le connette, come si vedrà, all'identità formale più profonda del confine sintassi-semantica: il gap informativo del capitolo III ha esattamente la struttura di un'impossibilità dimensionale.

questo capitolo sviluppa la famiglia II con la profondità che merita, costruendo i suoi strumenti formali principali — il teorema di Johnson-Lindenstrauss, la geometria degli spazi metrici, il duale della dimensione di Vapnik-Chervonenkis — e mostrando come essi unifichino risultati che i due cicli precedenti avevano trattato come separati.

## II. Il teorema di Johnson-Lindenstrauss: la matematica della compressione fedele

Il teorema di Johnson-Lindenstrauss (1984) è il risultato fondamentale della geometria della riduzione dimensionale. La sua formulazione più semplice afferma che un insieme di  $m$  punti in  $\mathbb{R}^D$  può essere proiettato in uno spazio di dimensione  $n = O(\log m / \epsilon^2)$  preservando tutte le distanze con distorsione al più  $\epsilon$ .

Teorema (Johnson-Lindenstrauss, 1984). Per ogni  $\epsilon \in (0,1)$  e ogni intero  $m > 0$ , sia  $n = \lceil C \epsilon^{-2} \log m \rceil$  per una costante universale  $C$ . Allora per ogni insieme di  $m$  punti  $x_1, \dots, x_m \in \mathbb{R}^D$ , esiste una mappa lineare  $f: \mathbb{R}^D \rightarrow \mathbb{R}^n$  tale che per tutti gli  $i, j \in \{1, \dots, m\}$ :

$$(1 - \epsilon) \|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1 + \epsilon) \|x_i - x_j\|^2$$

Informalmente: qualsiasi insieme finito di  $m$  punti può essere "compresso" in dimensione logaritmica preservando tutte le distanze a meno di un fattore  $(1 \pm \epsilon)$ . Il miracolo del teorema è che la dimensione target dipende soltanto da  $m$  (il numero di punti) e da  $\epsilon$  (la distorsione tollerata), non da  $D$  (la dimensione originale dello spazio). Un insieme di un milione di punti in dimensione un miliardo può essere compresso in dimensione  $O(\log 10^6 / \epsilon^2) \approx O(20/\epsilon^2)$  con distorsione  $\epsilon$ .

La dimostrazione standard usa proiezioni aleatorie: si sceglie una matrice  $R \in \mathbb{R}^{n \times D}$  con componenti indipendenti e identicamente distribuite secondo una

gaussiana standard  $N(0, 1/n)$  (o, equivalentemente, vettori uniformemente distribuiti sulla sfera unitaria in  $\mathbb{R}^D$ ). La mappa  $f(x) = (1/\sqrt{n}) R x$  è la proiezione aleatoria. Per ogni coppia  $(i, j)$ :

$$E[\|f(x_i) - f(x_j)\|^2] = \|x_i - x_j\|^2$$

e la varianza del quadrato della norma della proiezione è dell'ordine di  $\|x_i - x_j\|^4/n$ . L'argomento di concentrazione (la disuguaglianza di Johnson-Lindenstrauss originale, basata sul lemma di concentrazione delle misure di Lévy) mostra che con alta probabilità (probabilità  $\geq 1 - 2\exp(-\epsilon^2 n/8)$ ) la distanza proiettata è entro il fattore  $(1 \pm \epsilon)$  dalla distanza originale. Un union bound su tutte le  $C(m, 2)$  coppie di punti dà la probabilità di successo per l'intero insieme.

Il contenuto profondo del teorema non è il metodo della proiezione aleatoria — è il fatto che la dimensione logaritmica è sufficiente. Questo fatto ha un'interpretazione precisa: un insieme di  $m$  punti in posizione generica in  $\mathbb{R}^D$  ha "intrinsecamente" soltanto  $O(\log m)$  gradi di libertà geometrici rilevanti per le distanze mutue, indipendentemente dalla dimensione ambientale  $D$ . La "complessità geometrica" dell'insieme non è misurata dalla dimensione dello spazio ambiente ma dal numero di punti.

### III. Il converse di Johnson-Lindenstrauss: quando la compressione è impossibile

Il teorema di Johnson-Lindenstrauss afferma che la compressione fedele è possibile quando la dimensione target è almeno  $O(\log m / \epsilon^2)$ . Il suo converse — il risultato che stabilisce quando la compressione è impossibile — è altrettanto importante per le impossibilità dimensionali.

Teorema (lower bound di Johnson-Lindenstrauss, Alon 2003; Jayram-Woodruff 2013). Per ogni  $\epsilon \in (0, 1/2)$  e ogni intero  $m > 0$ , qualsiasi mappa  $f: \mathbb{R}^D \rightarrow \mathbb{R}^n$  che preserva tutte le distanze con distorsione  $(1 \pm \epsilon)$  per un insieme specifico di  $m$  punti richiede:

$$n \geq \Omega(\log m / \epsilon^2 \log(1/\epsilon))$$

In altri termini: la dimensione logaritmica non è soltanto sufficiente, è necessaria. Non si può fare meglio.

Questo lower bound ha una dimostrazione che usa la struttura combinatoriale degli insiemi di punti in posizione generale. Si costruisce un insieme di  $m$  punti in  $\mathbb{R}^D$  con la proprietà che qualsiasi mappa che distorce le distanze di meno di  $\epsilon$  deve distinguere tra  $\Omega(\log m)$  "direzioni" indipendenti — il che richiede dimensione target almeno logaritmica.

La costruzione dell'insieme "difficile" è la seguente: si considera l'ipercubo  $\{0,1\}^k$  in  $\mathbb{R}^k$ , che ha  $m = 2^k$  vertici. La distanza euclidea tra due vertici  $x, y \in \{0,1\}^k$  è  $\|x-y\| = \sqrt{\sum_i (x_i - y_i)^2} = \sqrt{dH(x,y)}$  dove  $dH$  è la distanza di Hamming. Qualsiasi mappa che preserva le distanze con distorsione  $(1 \pm \epsilon)$  deve preservare la struttura metrica dell'ipercubo — ma l'ipercubo in dimensione  $k$  ha  $m = 2^k$  vertici e dimensione intrinseca  $k$ . La compressione fedele dell'ipercubo richiede quindi dimensione target almeno  $\Omega(k) = \Omega(\log m)$ .

Questo converse del teorema di Johnson-Lindenstrauss è la struttura matematica sottostante a tutte le impossibilità dimensionali biologiche: quando il problema ha  $m$  "pattern" genuinamente distinti da distinguere, e si dispone di  $n < \Omega(\log m / \epsilon^2)$  gradi di libertà rappresentativi, la compressione fedele è impossibile.

#### IV. La formula di Cover come caso speciale del converse di JL

La formula di Cover per la capacità di separazione lineare (discussa nel capitolo I del ciclo biologico) è un caso speciale del converse di Johnson-Lindenstrauss, visto nella prospettiva della separabilità invece che della preservazione delle distanze. Vale la pena dimostrare questa connessione esplicitamente, perché rivela la struttura unitaria delle impossibilità dimensionali biologiche.

Ricordiamo la formula di Cover: la probabilità che  $m$  punti in posizione generica in  $\mathbb{R}^n$  siano linearmente separabili con un'etichettatura casuale è:

$$C(m, n) = (1/2^{m-1}) \cdot \sum_{k=0}^n \binom{m-1}{k} C(m-1, k)$$

con la transizione di fase  $C(m, n) \rightarrow 0$  per  $m > 2n$ . La formula dice che separare linearmente più di  $2n$  pattern in  $\mathbb{R}^n$  è improbabile (nel senso che una frazione vanishing delle etichettature casuali è separabile).

Nel linguaggio di Johnson-Lindenstrauss, questo si riformula come segue: la classificazione binaria (separazione lineare) di  $m$  punti richiede una funzione  $f: \mathbb{R}^n \rightarrow \{-1, +1\}$  che sia consistente con le etichette. La distorsione qui non è metrica

(preservazione delle distanze) ma topologica (preservazione delle separazioni). La dimensione critica  $n = \Omega(\log m)$  del teorema di JL corrisponde, nella terminologia della VC-teoria, alla dimensione di Vapnik-Chervonenkis  $d_{VC} = n + 1$  del percettrone lineare.

La connessione formale è questa: per un classificatore lineare in  $\mathbb{R}^n$ , la dimensione VC è  $n + 1$  (Capitolo I del ciclo biologico). Il teorema fondamentale della PAC learning afferma che la complessità di campionamento necessaria per la generalizzazione è:

$$m(\epsilon, \delta) = \Theta(d_{VC} / \epsilon + \log(1/\delta) / \epsilon)$$

dove  $\epsilon$  è l'errore tollerato e  $\delta$  è la probabilità di fallimento. La dimensione di VC  $n + 1$  è la versione discreta della dimensione  $n$  del teorema di Johnson-Lindenstrauss: in entrambi i casi,  $n$  misura il numero di gradi di libertà del sistema rappresentativo, e  $m/n$  (il rapporto tra la complessità del problema e la capacità del sistema) determina la fattibilità della rappresentazione.

La formula di Cover è la specializzazione della transizione di fase di JL al caso della classificazione: la soglia  $m = 2n$  della formula di Cover corrisponde alla soglia  $n = \Omega(\log m / \epsilon^2)$  del teorema di JL quando  $\epsilon = O(1)$  (distorsione dell'ordine dell'unità). Entrambe le soglie delimitano il regime in cui la rappresentazione è possibile da quello in cui è impossibile — la formula di Cover in termini di probabilità di separabilità, il teorema di JL in termini di fedeltà della proiezione.

## V. La dimensione intrinseca dei dati naturali

Il teorema di Johnson-Lindenstrauss afferma che qualsiasi insieme di  $m$  punti ha dimensione intrinseca  $O(\log m)$ . Ma gli insiemi di dati naturali — immagini, testi, segnali biologici — non sono insiemi generici in  $\mathbb{R}^D$ : sono concentrati su varietà di dimensione molto più bassa della dimensione ambientale. Questo è il contenuto dell'ipotesi della varietà (manifold hypothesis) discussa nel capitolo I del ciclo biologico, ma che ora possiamo sviluppare con maggiore profondità tecnica.

Sia  $X$  un insieme di dati con distribuzione  $P$  concentrata in un intorno di raggio  $\epsilon$  di una varietà  $\mathcal{M} \subset \mathbb{R}^D$  di dimensione  $d \ll D$ . Il teorema fondamentale della dimensione intrinseca (Whitney, 1944, nella versione moderna di Niyogi, Smale e

Weinberger, 2008) afferma che la varietà  $\mathcal{M}$  può essere ricostruita con alta probabilità da  $m$  punti campionati dalla distribuzione  $P$  se:

$$m \geq C \cdot (d / \epsilon^2) \cdot \log(d / \delta)$$

dove  $C$  è una costante universale e  $\delta$  è la probabilità di fallimento. Questo bound dipende da  $d$  (la dimensione intrinseca della varietà), non da  $D$  (la dimensione ambientale) — la ricostruzione è possibile con un numero di campioni logaritmico nella difficoltà e lineare nella dimensione intrinseca.

La dimensione intrinseca di diversi tipi di dati naturali è stimabile empiricamente. Per le immagini naturali di  $256 \times 256$  pixel (dimensione ambientale  $D = 65.536$ ), la dimensione intrinseca stimata varia tra 10 e 200 a seconda del tipo di immagine e del metodo di stima. Per le sequenze linguistiche (rappresentate come vettori in  $\mathbb{R}^{|V|}$  dove  $|V|$  è la dimensione del vocabolario, tipicamente  $|V| \approx 10^5$ ), la dimensione intrinseca stimata è dell'ordine di  $10^3$ – $10^4$ . Per i profili di espressione genica in cellule singole ( $D =$  numero di geni  $\approx 25.000$ ), la dimensione intrinseca stimata è dell'ordine di 10–50 tipi cellulari.

Questi valori mostrano che i dati naturali hanno struttura di varietà — sono concentrati su sottospazi di dimensione molto inferiore alla dimensione ambientale. Questa struttura è precisamente ciò che rende la cognizione biologica possibile: un cervello che rappresentasse ogni dimensione dello spazio sensoriale sarebbe impossibile con le risorse metaboliche disponibili. La struttura di varietà dei dati naturali riduce il problema della rappresentazione da  $D \approx 10^6$ – $10^9$  (la dimensione dello spazio di input sensoriale) a  $d \approx 10$ – $10^4$  (la dimensione intrinseca dello spazio dei concetti rilevanti) — una riduzione di diversi ordini di grandezza che rende fattibile la cognizione biologica.

Eppure questa struttura di varietà non elimina le impossibilità dimensionali — le sposta. Le impossibilità non riguardano più la rappresentazione di  $D$ -dimensioni in  $n$ -dimensioni, ma la rappresentazione di  $d$ -dimensioni in  $n$ -dimensioni, dove  $n$  è tipicamente comparabile con  $d$  o di poco superiore. Il confine tra rappresentazione fedele e rappresentazione con errore non scompare — si assottiglia, ma rimane.

## VI. Il tradeoff compressione-distorsione e il suo ruolo biologico

Le impossibilità dimensionali hanno una struttura di tradeoff che le distingue dalle altre famiglie: non vi è una soglia binaria tra "possibile" e "impossibile", ma una frontiera continua nel piano  $(n, \epsilon)$  che separa le coppie (dimensione target, distorsione tollerata) raggiungibili da quelle irraggiungibili.

Per il teorema di Johnson-Lindenstrauss, la frontiera è:

$$n \cdot \epsilon^2 = \Omega(\log m)$$

Le coppie  $(n, \epsilon)$  che soddisfano questa relazione sono raggiungibili; quelle che la violano no. La frontiera può essere attraversata in due direzioni: aumentando  $n$  (più dimensioni) a  $\epsilon$  fissato, oppure aumentando  $\epsilon$  (più distorsione) a  $n$  fissato. Il sistema cognitivo biologico naviga questa frontiera scegliendo implicitamente un punto nel piano  $(n, \epsilon)$  in base alle risorse disponibili ( $n$  è limitato dal numero di neuroni) e alla funzione richiesta ( $\epsilon$  è limitato dalla precisione necessaria per il comportamento adattivo).

Questa navigazione della frontiera ha una struttura evolutivamente precisa. In popolazioni diverse con pressioni selettive diverse, la posizione ottimale sul piano  $(n, \epsilon)$  è diversa: un predatore che deve identificare velocemente i movimenti delle prede ha bisogno di bassa latenza ( $n$  piccolo, elaborazione rapida) anche a costo di alta distorsione ( $\epsilon$  grande); un animale che dipende dalla discriminazione fine di dettagli morfologici (un uccello che distingue specie diverse per identificare i partner sessuali) ha bisogno di bassa distorsione ( $\epsilon$  piccolo) anche a costo di alta dimensionalità ( $n$  grande).

La diversità morfologica dei sistemi sensoriali biologici — la straordinaria varietà di tipi di fotorecettori, di meccanocettori, di chemiorecettori — è in parte una risposta evolutiva alla diversità dei punti ottimali sul piano  $(n, \epsilon)$  per le diverse pressioni selettive. Ogni architettura sensoriale è un punto sulla frontiera di Johnson-Lindenstrauss, scelto evolutivamente in funzione delle esigenze specifiche della nicchia ecologica.

## VII. La teoria delle $\epsilon$ -reti e la copertura degli spazi metrici

La struttura quantitativa delle impossibilità dimensionali si formalizza più completamente attraverso la teoria delle  $\epsilon$ -reti degli spazi metrici — un approccio che generalizza sia il teorema di Johnson-Lindenstrauss sia la formula di Cover in un framework unificato.

Sia  $(X, d)$  uno spazio metrico. Una  $\epsilon$ -rete di  $X$  è un sottoinsieme finito  $N_\epsilon \subset X$  tale che per ogni  $x \in X$  esiste  $n \in N_\epsilon$  con  $d(x, n) \leq \epsilon$ . Il numero di copertura (covering number)  $N(X, d, \epsilon)$  è la minima cardinalità di una  $\epsilon$ -rete di  $X$ .

Per gli spazi euclidei, il numero di copertura scala come:

$$N(B^d(r), \|\cdot\|, \epsilon) \approx (r/\epsilon)^d$$

dove  $B^d(r)$  è la palla di raggio  $r$  in  $\mathbb{R}^d$ . Questo mostra che la "complessità" di uno spazio euclideo di dimensione  $d$  è esponenziale in  $d$ : per coprire  $B^d(r)$  con palle di raggio  $\epsilon$ , servono dell'ordine di  $(r/\epsilon)^d$  palline.

Il rilevante per le impossibilità dimensionali è il numero di metrica (metric entropy)  $\log N(X, d, \epsilon)$ , che misura l'informazione necessaria per specificare un punto di  $X$  a precisione  $\epsilon$ . Per  $X = B^d(r)$ :

$$\log N(B^d(r), \|\cdot\|, \epsilon) = d \cdot \log(r/\epsilon) + O(d)$$

Il logaritmo del numero di copertura scala linearmente con la dimensione  $d$  — e questo è esattamente la misura della "difficoltà" di rappresentare uno spazio  $d$ -dimensionale con  $n < d$  gradi di libertà.

Il tradeoff compressione-distorsione del framework di Johnson-Lindenstrauss può essere riformulato in termini di numeri di copertura:

Lemma (riformulazione JL). Una mappa  $f: \mathbb{R}^D \rightarrow \mathbb{R}^n$  che preserva le distanze con distorsione  $(1 \pm \epsilon)$  può mappare al più  $N(B^n(r), \|\cdot\|, \epsilon) \approx (r/\epsilon)^n$  punti della palla  $B^D(r)$  con distorsione  $\epsilon$ . Per preservare  $m$  punti, serve  $n \geq \log m / \log(r/\epsilon) = \Omega(\log m)$  per  $\epsilon = O(1)$ .

Questo lemma mostra che il limite dimensionale del teorema di JL è un limite di entropia: lo spazio di rappresentazione  $\mathbb{R}^n$  ha entropia  $n \cdot \log(r/\epsilon)$ , e questa entropia deve essere sufficiente a distinguere tra  $m$  punti, il che richiede  $n \cdot \log(r/\epsilon) \geq \log m$ .

La versione biologica di questo lemma è diretta: la corteccia sensoriale con  $n$  neuroni ha capacità di rappresentare dell'ordine di  $(r/\epsilon)^n$  pattern distinti con precisione  $\epsilon$ . Per  $n = 10^6$  (un milione di neuroni corticali in un'area sensoriale) e  $\epsilon = 0.1$  (distorsione del 10%):

$$N \approx 10^{(10^6 \times \log_{10}(10))} = 10^{(10^6)}$$

un numero astronomico. Ma questa enorme capacità di rappresentazione si riduce drasticamente quando si considera la sparsità dell'attivazione: se soltanto il  $\varphi \approx 1\%$  dei neuroni è attivo per ogni stimolo, il numero di pattern distinti con attività sparsa è dell'ordine di:

$$C(10^6, 10^4) \approx 10^{(40.000)}$$

ancora molto grande, ma ridotto rispetto alla capacità del caso denso. La sparsità — la soluzione biologica alle impossibilità dimensionali discussa nel capitolo I del ciclo biologico — riduce l'entropia del sistema ma preserva la capacità di distinguere un numero molto elevato di pattern.

### VIII. Il teorema di isoperimetria e la concentrazione della misura

Un risultato matematico profondamente connesso alle impossibilità dimensionali è il fenomeno della concentrazione della misura negli spazi ad alta dimensione — il fatto controintuitivo che in spazi di dimensione alta, quasi tutta la "massa" di una distribuzione di probabilità è concentrata in un sottile guscio vicino alla frontiera dello spazio.

Il teorema isoperimetrico di Lévy-Milman afferma che per la sfera unitaria  $S^{(n-1)} \subset \mathbb{R}^n$  con la misura di Hausdorff normalizzata  $\sigma$ , e per qualsiasi insieme misurabile  $A \subset S^{(n-1)}$  con  $\sigma(A) \geq 1/2$ , l'espansione  $\epsilon$  di  $A$ :

$$A_\epsilon = \{x \in S^{(n-1)} : d(x, A) \leq \epsilon\}$$

ha misura

$$\sigma(A_\epsilon) \geq 1 - \exp(-(n-1)\epsilon^2/2)$$

Il complemento  $A_\epsilon^c$  ha misura esponenzialmente piccola per qualsiasi  $\epsilon > 0$  fissato e  $n \rightarrow \infty$ . La concentrazione della misura dice: in alta dimensione, quasi ogni punto è vicino a qualsiasi insieme che abbia misura almeno  $1/2$ .

La concentrazione della misura è il meccanismo che rende le proiezioni aleatorie di Johnson-Lindenstrauss funzionare: la norma di un vettore aleatorio in alta dimensione è concentrata intorno al suo valore atteso ( $\sqrt{n}$ ) con fluttuazioni dell'ordine di  $O(1)$ , non di  $O(\sqrt{n})$  come ci si aspetterebbe in bassa dimensione. Questa concentrazione è precisamente ciò che permette di preservare le distanze con distorsione  $(1 \pm \epsilon)$  con alta probabilità.

Ma la concentrazione della misura ha anche implicazioni controintuitive per le impossibilità dimensionali biologiche. In alta dimensione, due vettori scelti casualmente sono quasi ortogonali con probabilità che cresce esponenzialmente con la dimensione:

$$P(|\langle u, v \rangle| \leq \epsilon) \geq 1 - 2\exp(-n\epsilon^2/2) \text{ per } u, v \text{ sulla sfera unitaria}$$

Questo è il meccanismo della codifica sparsa: in alta dimensione, pattern diversi tendono naturalmente ad essere quasi ortogonali, riducendo le interferenze nel binding. La corteccia, operando in uno spazio di attivazione di dimensione  $n \approx 10^4$ – $10^6$ , beneficia della concentrazione della misura per mantenere basse le interferenze tra pattern diversi.

La connessione con il confine sintassi-semantica del capitolo III è ora visibile: la concentrazione della misura è il fenomeno che permette alle strutture sintattiche (i pattern di attivazione) di preservare approssimativamente le strutture semantiche (le relazioni tra concetti) in spazi di dimensione alta. La "magia" della cognizione in alta dimensione — il fatto che  $10^6$  neuroni possano rappresentare più concetti di quanti il numero di neuroni suggerirebbe — è una conseguenza della geometria degli spazi ad alta dimensione, non una proprietà speciale dei neuroni.

## IX. La dimensione di Hausdorff e le impossibilità frattali

Le impossibilità dimensionali che abbiamo sviluppato finora trattano la dimensione come un intero — il numero di coordinate indipendenti di uno spazio euclideo. Ma la dimensione, nella sua forma più generale, è un numero reale — la dimensione di Hausdorff — e le impossibilità dimensionali nei sistemi biologici spesso coinvolgono dimensioni frazionarie.

La dimensione di Hausdorff di un insieme  $A \subseteq \mathbb{R}^D$  è definita come:

$$\dim_H(A) = \inf\{d : H^d(A) = 0\} = \sup\{d : H^d(A) = \infty\}$$

dove  $H^d$  è la misura di Hausdorff d-dimensionale:

$$H^d(A) = \lim_{\epsilon \rightarrow 0} \inf \left\{ \sum_i (\text{diam } U_i)^d : A \subseteq \bigcup_i U_i, \text{diam } U_i \leq \epsilon \right\}$$

Per insiemi "normali" (curve differenziabili, superfici lisce),  $\dim_H$  coincide con la dimensione topologica. Per frattali — insiemi con struttura self-simile —  $\dim_H$  è un numero frazionario che cattura la complessità geometrica dell'insieme a scale diverse.

La dimensione frattale è rilevante per le impossibilità biologiche perché molte strutture naturali — la superficie polmonare, i rami vascolari, i dendriti neurali, le traiettorie del comportamento animale — hanno struttura frattale con dimensione frazionaria. Le impossibilità dimensionali per questi sistemi si formulano in termini di dimensione di Hausdorff invece che di dimensione euclidea.

Per un dendrite neurale con dimensione di Hausdorff  $\dim_H(D) = \delta \in (1, 3)$ , il numero di sinapsi che può ospitare nelle cellule di raggio  $\epsilon$  è dell'ordine di  $\epsilon^{-\delta}$  — un numero che cresce più rapidamente di  $\epsilon^{-1}$  (curva liscia) ma più lentamente di  $\epsilon^{-3}$  (volume euclideo). La capacità sinaptica del dendrite è quindi:

$$N_{\text{sinapsi}}(\epsilon) \approx C \cdot \epsilon^{-\delta}$$

con  $\delta \approx 1.5\text{--}2.0$  per i dendriti tipici. Questa struttura frattale ottimizza il numero di sinapsi per unità di volume metabolico — una soluzione biologica al tradeoff tra capacità di connettività e costo metabolico del cablaggio.

Il rilevante per le impossibilità è che la dimensione frattale impone un limite alla compressione: una struttura con  $\dim_H = \delta$  non può essere rappresentata fedelmente in una dimensione  $n < \delta$  — la sua struttura geometrica richiederebbe almeno  $\delta$  gradi di libertà. Per i dendriti con  $\delta \approx 1.8$ , la rappresentazione in dimensione 1 (linea) o dimensione 2 (superficie) sarebbe necessariamente approssimata, con errori dell'ordine  $\epsilon^{\{\delta - n\}}$  per ogni scala  $\epsilon$ .

## **X. Le impossibilità dimensionali come oracoli biologici**

La conclusione che questo capitolo intende sostenere è che le impossibilità dimensionali non sono soltanto limiti — sono oracoli biologici. Dicono ai sistemi biologici dove concentrare la propria complessità, in quale direzione investire le

risorse evolutive, quale struttura adottare per massimizzare la funzione cognitiva entro i vincoli fisici disponibili.

Il teorema di Johnson-Lindenstrauss dice: se la dimensione target è  $n$ , si possono separare fedelmente fino a  $2^{\Omega(n)}$  punti. La selezione evolutiva per la massima capacità cognitiva entro il vincolo  $n = O(\text{neuroni disponibili})$  è quindi la selezione per la massima esplorazione dello spazio degli esponenziali — la selezione per la codifica sparsa, per la struttura di varietà, per la concentrazione della misura.

Il converse di Johnson-Lindenstrauss dice: non si può fare meglio. Qualsiasi architettura cognitiva alternativa che tentasse di superare il bound logaritmico incontrerebbe necessariamente impossibilità dimensionali — distorsioni, interferenze, fallimenti del binding. L'architettura biologica osservata — sparsità, varietà, gerarchia — non è una delle tante soluzioni possibili: è la famiglia di soluzioni ottimali alla frontiera di Johnson-Lindenstrauss.

Questo è il senso in cui le impossibilità dimensionali sono oracoli: forniscono informazione normativa sulla struttura ottimale dei sistemi cognitivi. Non dicono soltanto cosa non si può fare — dicono cosa si deve fare per stare il più vicino possibile al confine della possibilità. E il cervello biologico, risultato di miliardi di anni di ottimizzazione evolutiva, si trova esattamente su quel confine — non per caso, ma perché è l'unico punto dove la cognizione reale è possibile.

---

## Capitolo V

### Le impossibilità di complessità come geometria dello spazio dei problemi

#### La struttura della difficoltà computazionale, la gerarchia polinomiale e il suo significato epistemologico

##### Mappa epistemica — Capitolo V

| Status          | Cod. | Claim                                                                                                                                                       |
|-----------------|------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | La gerarchia polinomiale (Stockmeyer 1976; Fagin 1974) e la $\Pi_2^P$ -completezza della Circumscription (Liberatore-Schaerf 1998): formalmente dimostrati. |
| CONGETTURA      | C    | $P \neq NP$ : non dimostrata. Tutte le impossibilità della Famiglia III dipendono da questa congettura.                                                     |
| CONGETTURA      | C    | Baker-Gill-Solovay (1975): nessun metodo relativizzante può risolvere $P$ vs $NP$ , suggerendo fortemente $P \neq NP$ ma non dimostrandolo.                 |
| INTERPRETAZIONE | I    | La creatività umana come asimmetria strutturale trovare/verificare (Aaronson 2009): proposta interpretativa potente, non derivazione formale.               |

#### I. Il problema che non è impossibile ma è inaccessibile

Esiste una categoria di limite intellettuale che non appartiene né all'impossibilità logica assoluta né alla mera difficoltà pratica: la classe dei problemi per cui esistono algoritmi risolutori in linea di principio, ma per cui qualsiasi algoritmo richiede risorse che eccedono quelle disponibili a qualsiasi sistema fisicamente realizzabile in ogni tempo utile. Un problema di questa categoria non è "impossibile" nel senso in cui la quadratura del cerchio è impossibile — con risorse infinite potrebbe essere risolto. Ma con le risorse fisiche dell'universo reale — con il tempo rimasto prima della morte termica, con l'energia disponibile in qualsiasi struttura materiale finita — è funzionalmente inaccessibile.

La teoria della complessità computazionale è la matematica di questa categoria — la disciplina che classifica i problemi non in base alla loro risolubilità ma in base alle risorse necessarie per risolverli. E la sua tesi fondamentale — quella che, se

dimostrata, costituirebbe uno dei risultati matematici più profondi della storia — è che la gerarchia delle difficoltà è strettamente crescente: esistono problemi intrinsecamente più difficili di altri, non per incapacità degli algoritmi attuali ma per proprietà strutturali dei problemi stessi.

questo capitolo propone una lettura inedita della teoria della complessità computazionale: non soltanto come classificazione delle risorse necessarie, ma come geometria dello spazio dei problemi. La gerarchia polinomiale  $P \subseteq NP \subseteq PH \subseteq PSPACE \subseteq EXPTIME$  non è soltanto una scala di difficoltà crescente — è una struttura topologica sullo spazio dei problemi computazionali, e la posizione di un problema in questa struttura dice qualcosa di profondo sulla natura del problema, non soltanto sulla difficoltà di risolverlo. In questa lettura, le impossibilità di complessità si connettono alle altre famiglie di impossibilità strutturali in modi che la teoria standard della complessità non cattura.

## II. La struttura della gerarchia polinomiale

La gerarchia polinomiale (PH) è costruita iterativamente a partire dalla classe P (problemi risolvibili in tempo polinomiale da una macchina di Turing deterministica) attraverso l'operazione di aggiunta di un oracolo per i problemi del livello precedente.

Le classi fondamentali al primo livello sono:

$P = \{L : \exists \text{ macchina di Turing } M \text{ deterministica che decide } L \text{ in tempo } p(n) \text{ per un polinomio } p\}$

$NP = \{L : \exists \text{ macchina di Turing } M \text{ non-deterministica che decide } L \text{ in tempo } p(n)\} = \{L : \exists \text{ verificatore } V \text{ e un polinomio } p \text{ tali che } x \in L \leftrightarrow \exists w, |w| \leq p(|x|), V(x, w) = 1\}$

$co-NP = \{L : L^c \in NP\}$

Il carattere fondamentale di NP è che i problemi in NP sono quelli per cui una soluzione può essere verificata in tempo polinomiale, anche se trovarla può richiedere tempo esponenziale. L'esistenzialità è la struttura chiave:  $x \in L \leftrightarrow \exists w, V(x, w) = 1$ . La struttura di co-NP è invece quella della universalità:  $x \in L^c \leftrightarrow \forall w, V(x, w) = 0$ , ovvero  $x \in L \leftrightarrow \forall w, V(x, w) = 1$  nella versione negata.

Al secondo livello:

$\Sigma_2^P = NP^{\{NP\}} = \{L : L \text{ è decidibile in tempo polinomiale da una macchina non-deterministica con accesso a un oracolo } NP\}$

$\Pi_2^P = co-NP^{\{NP\}} = \text{complemento di } \Sigma_2^P$

La caratterizzazione alternativa di  $\Sigma_2^P$  è:  $L \in \Sigma_2^P \leftrightarrow \exists$  polinomio  $p$  e predicato  $R$  calcolabile in polinomio tale che:

$$x \in L \leftrightarrow \exists y, |y| \leq p(|x|), \forall z, |z| \leq p(|x|), R(x, y, z) = 1$$

La struttura è  $\exists \forall$  — un quantificatore esistenziale seguito da uno universale.  $\Pi_2^P$  ha la struttura  $\forall \exists$ . Il problema del frame come Circumscription — decidere se una formula è una conseguenza della minimizzazione — ha esattamente la struttura  $\forall \exists$ : per ogni modello  $M'$  che soddisfa l'insieme di formule con un numero di atomi cambiati non superiore alla minimizzazione, vale la conseguenza? La struttura universale sull'insieme dei modelli seguita dalla struttura esistenziale sulle conseguenze è la firma della  $\Pi_2^P$ -completezza.

Genericamente, il  $k$ -esimo livello della gerarchia polinomiale è:

$\Sigma^k P = \{L : L \in (\Sigma^{k-1} P)^{\{NP\}}\}$  —  $k$  quantificatori alternanti, partendo da  $\exists$

$\Pi^k P = co-\Sigma^k P$  —  $k$  quantificatori alternanti, partendo da  $\forall$

$$PH = \bigcup \Sigma^k P = \bigcup \Pi^k P$$

Teorema (Stockmeyer, 1976). Se  $PH$  collassa al  $k$ -esimo livello ( $\Sigma^k P = \Pi^k P$ ), allora  $PH = \Sigma^k P$ . In particolare, se  $P = NP$ , allora  $PH = P$ .

Il risultato di Stockmeyer mostra che la gerarchia è robusta: se  $P \neq NP$  (la congettura dominante), la gerarchia non collassa e tutti i livelli sono genuinamente distinti. Ogni problema che è  $\Sigma^k P$ -completo è strettamente più difficile di qualsiasi problema in  $\Sigma^{k-1} P$ .

### III. La geometria della riduzione: lo spazio dei problemi come struttura di ordine parziale

La teoria della complessità classifica i problemi attraverso le riduzioni — le trasformazioni computazionalmente efficienti tra problemi. Il problema  $A$  si riduce al problema  $B$  ( $A \leq_m P B$ , riduzione many-one in tempo polinomiale) se

esiste una funzione  $f$  calcolabile in tempo polinomiale tale che  $x \in A \leftrightarrow f(x) \in B$ . Se  $A \leq_m^P B$  e  $B \in C$ , allora  $A \in C$  (le riduzioni "trainano" i problemi verso classi più facili).

Questa struttura di riduzioni definisce un ordine parziale sullo spazio dei problemi:  $A \leq_m^P B$  significa che  $B$  è "almeno difficile quanto"  $A$ . I problemi  $C$ -completi per una classe  $C$  sono i massimali in questo ordine parziale: i problemi più difficili della classe  $C$ , a cui tutti gli altri problemi di  $C$  si riducono. La  $C$ -completezza è una proprietà strutturale del problema, non dell'algoritmo.

La geometria di questo ordine parziale ha proprietà notevoli. Per la classe NP:

- Esiste un numero astronomico di problemi NP-completi — problemi in matematica discreta, in ottimizzazione combinatoriale, in teoria dei grafi, in logica, in bioinformatica, in fisica statistica. Tutti appartengono allo stesso punto massimale dell'ordine parziale NP.
- Questo punto massimale non è raggiunto da problemi "patologici" o costruiti ad arte: i problemi NP-completi includono problemi di grandissimo interesse pratico (la soddisfacibilità booleana 3-SAT, il problema del commesso viaggiatore, la colorazione dei grafi, il problema dello zaino, la pianificazione logistica).
- Tra  $P$  e NP, se  $P \neq NP$ , esistono problemi NP-intermedi — problemi che sono in NP ma non sono né in  $P$  né NP-completi (teorema di Ladner, 1975). Il fattorizzazione degli interi è un candidato congetturato per questa classe.

La struttura geometrica dell'ordine parziale è quindi: un punto minimo  $P$  (i problemi facili), un punto massimo NP-completo (i problemi più difficili di NP), e una regione intermedia NP-intermedia di struttura non ancora completamente caratterizzata. Questa non è una struttura lineare — è una struttura reticolare complessa in cui la posizione di ogni problema riflette le sue proprietà strutturali intrinseche.

#### IV. Perché $P \neq NP$ sarebbe una impossibilità strutturale e non tecnica

La congettura  $P \neq NP$ , se vera, è l'affermazione che esistono problemi in NP che richiedono tempo superpolinomiale per qualsiasi algoritmo deterministico. Ma il significato epistemologico di questa affermazione va molto al di là della teoria della complessità — è un'affermazione sulla struttura cognitiva del pensiero creativo.

I problemi in NP hanno la struttura dell'invenzione e della verifica: trovare una soluzione può essere difficile (il tempo esponenziale della ricerca nello spazio delle soluzioni), ma verificare che una soluzione proposta sia corretta è facile (il tempo polinomiale del verificatore). La domanda  $P = NP?$  chiede: può la ricerca essere altrettanto efficiente della verifica?

La risposta  $NP \neq P$ , se vera, afferma: no. Trovare è strutturalmente più difficile di verificare. Questo non è soltanto un fatto computazionale — è un fatto epistemologico: la creazione è strutturalmente diversa dal riconoscimento. Una prova matematica può essere verificata in tempo ragionevole, ma trovare quella prova richiede un processo creativo che non è sistematizzabile in un algoritmo efficiente. Un'opera d'arte può essere giudicata (verificata come esteticamente riuscita) molto più facilmente di quanto possa essere creata (prodotta come esito di una ricerca nello spazio delle composizioni possibili).

Scott Aaronson (2009) ha articolato questa connessione tra  $P \neq NP$  e la "specialità della creatività umana" con la seguente osservazione: se  $P = NP$ , allora chiunque possa apprezzare una bella dimostrazione matematica potrebbe anche produrla — il giudizio e la produzione sarebbero computazionalmente equivalenti. Se  $P \neq NP$ , apprezzare è più facile di produrre, e questa asimmetria è strutturale — non un limite accidentale della nostra intelligenza, ma una proprietà del problema stesso.

La formalizzazione di questa asimmetria usa la struttura dei quantificatori: i problemi di verifica hanno la struttura  $\exists w, V(x, w) = 1$  (esiste una soluzione? — verificare richiede soltanto di controllare un testimone). I problemi di ricerca hanno la struttura di trovare  $w$  tale che  $V(x, w) = 1$  (trovare la soluzione — navigare lo spazio dei testimoni). Se  $P \neq NP$ , la struttura  $\exists$  non è computazionalmente equivalente alla struttura di ricerca — esistere non implica essere trovabile efficientemente.

Questo è il contenuto epistemologico della congettura: la creatività — la ricerca di soluzioni a problemi per cui la verifica è facile — è strutturalmente irriducibile alla verifica. Non è un caso che i sistemi cognitivi biologici abbiano evoluto meccanismi separati per la ricerca creativa (il sistema default mode, l'incubazione, l'insight) e per la verifica (il sistema di controllo esecutivo, il ragionamento deduttivo): la struttura della complessità computazionale suggerisce che questi processi debbano essere strutturalmente distinti.

## V. La $\Pi_2^P$ -completezza del frame problem: analisi profonda

La  $\Pi_2^P$ -completezza della Circumscription di McCarthy, dimostrata da Liberatore e Schaerf (1998), è il risultato più direttamente rilevante per la biologia cognitiva tra quelli della teoria della complessità. Ma la sua dimostrazione rivela una struttura che merita un'analisi più approfondita di quanto i cicli precedenti abbiano condotto.

La Circumscription di McCarthy è una forma di ragionamento per chiusura minima: tra tutti i modelli che soddisfano un insieme di formule  $F$ , si preferiscono i modelli che minimizzano l'estensione di certi predicati (tipicamente i predicati che descrivono le proprietà che cambiano). La preferenza per i modelli minimali cattura l'intuizione che, in assenza di informazione esplicita su un cambiamento, si assume che il cambiamento non sia avvenuto.

Formalmente, dato un insieme di formule  $F$  in logica del primo ordine e un predicato  $P$  da minimizzare, il modello di Circumscription  $CIRC(F; P)$  è il modello  $M$  di  $F$  tale che non esiste  $M' \models F$  con  $P^{\{M'\}} \subsetneq P^M$  (l'estensione di  $P$  in  $M'$  è strettamente contenuta in quella di  $M$ ). La conseguenza logica nella Circumscription è:

$F \models_{CIRC} \varphi \leftrightarrow \varphi$  è vera in tutti i modelli minimali di  $F$

La struttura di questa definizione è  $\forall \exists$ : per ogni modello  $M$  ( $\forall$ ), se  $M$  è minimale per  $P$  (una condizione che richiede di verificare  $\forall M'$ , se  $M' \models F$  allora  $P^M \subseteq P^{\{M'\}}$ , — un altro  $\forall$ ), allora  $\varphi$  è vera in  $M$ . La struttura complessiva è  $\forall M (\forall M', \text{cond}(M, M')) \rightarrow \varphi(M)$ , che è una formula  $\Pi_2$  (due quantificatori universali annidati, o equivalentemente  $\forall \exists$  con la condizione di minimalità che ha struttura  $\exists$ ).

La  $\Pi_2^P$ -completezza non è soltanto un risultato tecnico: rivela che il ragionamento per chiusura minima — il modo in cui il sistema cognitivo gestisce l'ignoranza sulle conseguenze delle azioni — ha struttura intrinsecamente più difficile di NP. Anche il tipo di ragionamento "cosa posso dedurre da quello che so?" è computazionalmente inaccessibile in generale, il che spiega perché i sistemi cognitivi biologici usano approssimazioni (i prior ecologici, le euristiche) invece dell'inferenza ottimale: l'inferenza ottimale è  $\Pi_2^P$ -completa, e nessun sistema biologico ha le risorse per realizzarla.

## VI. La relazione tra complessità e confine sintassi-semantica

il capitolo III ha argomentato che tutte le impossibilità strutturali hanno al proprio centro il confine sintassi-semantica. Come si inseriscono in questa struttura le impossibilità di complessità?

La connessione è meno diretta di quella per le famiglie I e IV, ma è reale e profonda. La struttura dei quantificatori della gerarchia polinomiale —  $\exists, \forall, \exists \forall, \forall \exists, \exists \forall \exists, \dots$  — è precisamente la struttura della semantica della logica del primo ordine. Le formule  $\Sigma_1^P$  hanno la forma  $\exists x, \varphi(x)$  — la forma esistenziale che è il fondamento della semantica del primo ordine. Le formule  $\Pi_1^P$  hanno la forma  $\forall x, \varphi(x)$  — la forma universale. I livelli superiori della gerarchia corrispondono a formule più profondamente quantificate — formule che fanno affermazioni semantiche su quantificatori di ordine superiore.

Il teorema di Fagin (1974) rende esplicita questa connessione: un problema è in NP se e soltanto se è definibile da una sentenza del secondo ordine esistenziale ( $\exists R, \varphi(R, x)$  dove  $R$  è un simbolo di relazione del secondo ordine e  $\varphi$  è del primo ordine). Questo è il teorema di caratterizzazione della complessità per via della logica:  $NP = \exists SO$  (le formule del secondo ordine con un quantificatore esistenziale sul simbolo relazionale di secondo ordine).

La gerarchia polinomiale corrisponde alla gerarchia delle formule del secondo ordine:

$\Sigma^P_k =$  formule del secondo ordine con  $k$  quantificatori alternanti, partendo da  $\exists$

$\Pi^P_k =$  formule del secondo ordine con  $k$  quantificatori alternanti, partendo da  $\forall$

PSPACE = formule del secondo ordine con quantificatori del secondo ordine universali

La connessione con il confine sintassi-semantica è ora visibile: le impossibilità di complessità sono le impossibilità di ragionare sulle strutture semantiche (i modelli, le interpretazioni, i quantificatori del secondo ordine) usando soltanto risorse sintattiche (le computazioni deterministiche in tempo polinomiale). La  $\Pi_2^P$ -completezza del frame problem significa che ragionare sulla semantica della chiusura minima richiede due livelli di quantificazione oltre il primo ordine — due livelli di distanza dalla sintassi verso la semantica.

Questo chiarisce il legame tra le impossibilità di complessità e il confine sintassi-semantica: le impossibilità di complessità misurano la profondità semantica di un problema — quanti livelli di quantificazione del secondo ordine sono necessari per caratterizzarlo. La congettura  $P \neq NP$  afferma che la profondità semantica 1 ( $\exists SO$ ) non è riducibile alla profondità 0 (logica del primo ordine con quantificatori soltanto sul dominio degli individui). Le gerarchie di complessità superiori affermano che le profondità semantiche superiori non sono riducibili alle inferiori.

## **VII. Il teorema di Baker-Gill-Solovay e i limiti dei metodi di dimostrazione**

Il risultato di Baker, Gill e Solovay (1975) è uno dei più inquietanti nella teoria della complessità: esistono oracoli  $A$  e  $B$  tali che  $P^A = NP^A$  e  $P^B \neq NP^B$ . In altre parole, rispetto ad alcuni oracoli,  $P$  e  $NP$  coincidono; rispetto ad altri, sono distinte. Questo implica che nessun metodo di dimostrazione che "relativizza" — che funziona ugualmente bene in presenza di oracoli — può risolvere la questione  $P$  vs  $NP$ .

La maggior parte dei metodi di dimostrazione della teoria della complessità relativizza: le diagonalizzazioni, le simulazioni, le riduzioni — tutte queste tecniche sono "indipendenti dall'oracolo" nel senso che funzionano identicamente in presenza o in assenza di un oracolo. Il teorema di Baker-Gill-Solovay mostra che questi metodi sono insufficienti: qualsiasi dimostrazione di  $P \neq NP$  deve usare tecniche non-relativizzanti.

Le tecniche non-relativizzanti identificate nella letteratura successiva sono:

La tecnica algebrica: i metodi che usano la struttura algebrica dei problemi (come la rappresentazione polinomiale dei circuiti booleani). Il teorema  $IP = PSPACE$  di Shamir (1992) usa questa tecnica — la prova interattiva di proprietà del polinomiale permanente.

La tecnica naturale delle prove: Razborov e Rudich (1997) hanno dimostrato che le "prove naturali" — una classe ampia di tecniche combinatoriali che includeva la maggior parte dei tentativi di lower bound — non possono dimostrare  $P \neq NP$  se certi oggetti crittografici esistono (specificamente, se esistono pseudogeneratori di bit). Le prove naturali sono esse stesse limitate da un'impossibilità strutturale.

Questo è un caso notevole di impossibilità di secondo ordine: non soltanto il problema  $P$  vs  $NP$  è difficile, ma la struttura delle prove che potrebbero risolverlo è soggetta a proprie impossibilità strutturali. La difficoltà del problema si riflette nella difficoltà di dimostrare la difficoltà del problema — una struttura ricorsiva che ricorda la costruzione diagonale del capitolo II.

Il significato epistemologico del teorema Baker-Gill-Solovay è profondo: la questione  $P$  vs  $NP$  è intrinsecamente non-locale — non può essere risolta da metodi che non tengono conto della struttura globale dei problemi computazionali. Questo è esattamente la struttura del confine sintassi-semantica: la semantica (il significato globale del problema) non è catturabile dalla sintassi (le manipolazioni locali degli algoritmi) senza risorse che eccedono il livello sintattico corrente.

## **VIII. La complessità della comunicazione e le impossibilità distribuite**

Le impossibilità di complessità non si limitano ai problemi risolti da un singolo agente computazionale. Esiste una ricca teoria della complessità della comunicazione — le risorse necessarie quando il problema deve essere risolto da più agenti con informazioni distribuite — che rivela impossibilità strutturali di tipo diverso, più direttamente rilevanti per le impossibilità biologiche nei sistemi sociali e cognitivi distribuiti.

Il framework della complessità della comunicazione, introdotto da Yao (1979), considera due agenti Alice e Bob che hanno rispettivamente gli input  $x$  e  $y$ , e devono computare una funzione  $f(x, y)$  comunicando il minor numero possibile di bit. La complessità della comunicazione  $CC(f)$  è il minimo numero di bit scambiati nel protocollo ottimale nel caso peggiore.

Un risultato fondamentale riguarda la funzione  $DISJ(x, y) = 1 \leftrightarrow x \cap y = \emptyset$  (la disgiunzione degli insiemi), dove  $x, y \subseteq \{1, \dots, n\}$ . La complessità della comunicazione di  $DISJ$  è  $\Theta(n)$  — Alice e Bob devono comunicare  $\Omega(n)$  bit nel caso peggiore per determinare se i loro insiemi sono disgiunti. Questo lower bound, dovuto a Kalyanasundaram e Schnitger (1992) e semplificato da Razborov (1992), usa il metodo degli elementi intersecanti: se la comunicazione fosse minore di  $n$  bit, gli agenti potrebbero comprimere le loro informazioni in meno bit di quanti ne servono per distinguere tra tutti i possibili input.

La versione multi-agente della complessità della comunicazione — con  $k$  agenti che comunicano su un grafo di comunicazione  $G$  — è ancora più rilevante per le impossibilità biologiche. Se il grafo  $G$  non è completamente connesso, esistono funzioni globali  $f(x_1, \dots, x_k)$  che richiedono una quantità di comunicazione proporzionale alla distanza di taglio (cut size) del grafo — la struttura topologica del grafo di comunicazione determina la complessità della comunicazione.

Applicato ai sistemi cognitivi distribuiti: la corteccia cerebrale (con la sua struttura small-world e il suo rich-club, analizzata nel capitolo X del ciclo biologico) è il grafo di comunicazione del sistema cognitivo distribuito che il cervello implementa. La complessità delle funzioni cognitive che il cervello può computare è limitata dalla struttura topologica di questo grafo — non soltanto dai cicli di clock dei neuroni o dalla loro densità. Le impossibilità di complessità nella cognizione sono impossibilità di complessità della comunicazione nel grafo del connettoma.

## IX. Il teorema PCP e la non-approssimabilità

Il teorema PCP (Probabilistically Checkable Proofs, Arora et al. 1998) è uno dei risultati più sorprendenti della teoria della complessità e ha implicazioni dirette per la comprensione delle impossibilità di approssimazione — una categoria di

impossibilità strettamente legata alle impossibilità biologiche di cognizione ottimale.

Teorema PCP (Arora e Safra, 1998; Arora et al., 1998).  $NP = PCP(O(\log n), O(1))$  — ogni problema in NP ha una prova "verificabile probabilisticamente" che richiede soltanto  $O(\log n)$  bit di accesso casuale alla prova e  $O(1)$  query alla prova, con errore al più  $1/2$ .

La conseguenza fondamentale del teorema PCP per l'approssimazione è che molti problemi di ottimizzazione NP-hard non sono approssimabili in modo arbitrariamente preciso in tempo polinomiale. Specificamente: se il problema MAX-3SAT (massimizzare il numero di clausole soddisfatte in una formula 3-CNF) potesse essere approssimato con fattore migliore di  $7/8$  in tempo polinomiale, allora  $P = NP$  (teorema di Håstad, 2001). Il fattore  $7/8$  è il "gap di approssimabilità" — la soglia al di sotto della quale l'approssimazione diventa computazionalmente impossibile.

Questo gap di approssimabilità ha una struttura formale precisa: è determinato dalla distanza tra il valore ottimale OPT della soluzione e il valore massimo raggiungibile da algoritmi polinomiali. Per MAX-3SAT, il gap è  $1 - 7/8 = 1/8$ : qualsiasi algoritmo polinomiale non può avvicinarsi più del 12.5% al valore ottimale nel caso peggiore.

La connessione con le impossibilità biologiche è la seguente: i sistemi cognitivi biologici, operando con risorse computazionali limitate, si trovano sistematicamente nella regione del gap di approssimabilità. Il cervello non ottimizza — approssima. E le sue approssimazioni (le euristiche, i bias, i prior ecologici) sono strategie che operano precisamente nel regime in cui l'ottimizzazione esatta è computazionalmente inaccessibile. Il fatto che il cervello usi euristiche invece dell'ottimizzazione bayesiana esatta non è una limitazione evolutiva contingente — è una risposta adattiva alle impossibilità di complessità che rendono l'ottimizzazione esatta inaccessibile in tempo reale.

Il teorema PCP, letto in questa chiave, afferma che la cognizione è necessariamente approssimata — non per limitazioni del substrato biologico, ma per impossibilità strutturali del tipo di problemi che la cognizione deve risolvere. L'architettura del cervello (con i suoi prior, le sue euristiche, i suoi errori sistematici) è

l'implementazione biologica di una struttura di approssimazione ottimale sotto i vincoli di complessità che il teorema PCP rivela come irriducibili.

## **X. La complessità come misura della profondità del problema: verso una tesi unitaria**

questo capitolo ha sviluppato la Famiglia III — le impossibilità di complessità — come una struttura geometrica sullo spazio dei problemi, non come una lista di risultati tecnici. La gerarchia polinomiale è una topologia sullo spazio dei problemi computazionali; le riduzioni sono le mappe che definiscono la struttura di ordine parziale; la completezza è la proprietà dei massimali di quest'ordine; il gap di approssimabilità è la misura quantitativa della distanza tra l'ottimo raggiungibile in tempo polinomiale e l'ottimo assoluto.

Ma la tesi più profonda che questo capitolo propone — e che le successive svilupperanno — è che le impossibilità di complessità non sono indipendenti dalle altre famiglie: sono il versante quantitativo del confine sintassi-semantica del capitolo III. La profondità di un problema nella gerarchia polinomiale misura quanti livelli semantici deve attraversare per essere definito — quanta struttura di secondo ordine è necessaria per caratterizzarlo. La congettura  $P \neq NP$  afferma che il primo livello semantico ( $\exists SO$ ) è irriducibile al livello sintattico (FO del primo ordine); le gerarchie superiori affermano che ogni livello semantico è irriducibile al precedente.

In questa lettura, le impossibilità di complessità non sono una famiglia separata dalle impossibilità di chiusura semantica — sono le impossibilità di chiusura semantica misurate in termini di risorse computazionali necessarie per attraversare il confine. Ogni volta che un problema richiede un livello superiore della gerarchia polinomiale, richiede un accesso supplementare alla struttura semantica del problema — un salto nella gerarchia tarskiana che il sistema computazionale deve eseguire.

La geometria dello spazio dei problemi è quindi anche una geometria del confine sintassi-semantica: la posizione di un problema nella gerarchia polinomiale misura la distanza tra la sua struttura sintattica (la stringa di input, la formula da soddisfare, il grafo da colorare) e la sua struttura semantica (la soluzione ottimale,

la verità della formula, la colorazione corretta). Le impossibilità di complessità sono, in ultima analisi, impossibilità di attraversare quella distanza con risorse fisicamente finite.

---

## Capitolo VI

### L'irreversibilità come quinta famiglia

La freccia del tempo nelle impossibilità strutturali, la path dependence formale e il teorema dell'entropia cognitiva

Mappa epistemica — Capitolo VI

| Status          | Cod. | Claim                                                                                                                                                                                   |
|-----------------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | Il principio di Landauer (1961) e la sua verifica sperimentale (Bérut et al. 2012): fisicamente fondato e verificato.                                                                   |
| DIMOSTRATO      | D    | La struttura formale dell'irreversibilità come morfismo non-iniettivo.                                                                                                                  |
| CONGETTURA      | C    | Il ratchet di Müller come istanza biologica del II principio nelle popolazioni genetiche: congettura ben fondata.                                                                       |
| INTERPRETAZIONE | I    | La chiusura dei periodi critici come transizione di fase irreversibile; la deriva epigenetica come versione cellulare di Landauer: proposte interpretative con forte supporto empirico. |

#### I. Una lacuna nella tassonomia

La tassonomia delle quattro famiglie proposta nel capitolo I di questo ciclo era consapevolmente provvisoria: si preannunciava che dall'analisi degli aspetti dinamici delle impossibilità strutturali sarebbe emersa una quinta famiglia che le altre quattro non catturano. questo capitolo mantiene quella promessa.

Le quattro famiglie precedenti condividono una caratteristica strutturale che si rivela soltanto quando si accosta questo ciclo ai cicli precedenti: sono tutte impossibilità sincroniche. Le impossibilità diagonali, dimensionali, di complessità e di chiusura semantica affermano che certe strutture non possono essere realizzate in un dato momento — non che certi processi non possano essere invertiti nel tempo. La famiglia dell'irreversibilità è qualitativamente diversa: afferma che certi processi, una volta avvenuti, non possono essere invertiti — che il tempo introduce una asimmetria strutturale che le impossibilità sincroniche non catturano.

L'irreversibilità come famiglia delle impossibilità strutturali abbraccia fenomeni molto diversi in superficie: la chiusura dei periodi critici (Capitolo V del ciclo biologico), il ratchet di Müller (Capitolo II del ciclo biologico), la deriva epigenetica (Capitolo XVI del ciclo biologico), il catastrophic forgetting nell'apprendimento (Capitolo V del ciclo biologico), la degradazione della trasmissione culturale (Capitolo VII del ciclo biologico), e persino la freccia del tempo termodinamica (Capitolo XII del ciclo biologico). questo capitolo dimostra che tutti questi fenomeni hanno la medesima struttura formale — la struttura dell'impossibilità di invertire una transizione in un sistema con struttura di stati asimmetrica — e che questa struttura è connessa alle quattro famiglie precedenti attraverso la relazione con il confine sintassi-semantica.

## II. La struttura formale dell'irreversibilità

L'irreversibilità è tipicamente definita in termodinamica attraverso il secondo principio: un processo è irreversibile se il suo inverso temporale non è un processo spontaneamente possibile per il sistema isolato. La formulazione matematica precisa usa la nozione di entropia di Shannon-Boltzmann.

Per un sistema con stati  $s \in S$  e distribuzione di probabilità  $p$ , l'entropia è:

$$H(p) = -\sum_s p(s) \log p(s)$$

Il secondo principio afferma che per un sistema isolato in evoluzione secondo una dinamica stocastica markoviana  $T: \Delta(S) \rightarrow \Delta(S)$  (dove  $\Delta(S)$  è lo spazio delle distribuzioni di probabilità su  $S$ ):

$$H(T(p)) \geq H(p)$$

L'entropia non decresce — il sistema evolve verso distribuzioni di maggiore disordine. La dinamica inversa  $T^{-1}$  (che farebbe diminuire l'entropia) è possibile in linea di principio ma richiede conoscenza perfetta dello stato microscopico corrente — conoscenza che un macrosistema non può avere in pratica. L'irreversibilità è quindi una conseguenza della perdita di informazione microscopica durante l'evoluzione macroscopica.

Ma questa è la forma termodinamica dell'irreversibilità, che è già stata analizzata nel capitolo XII del ciclo biologico. La forma che questo capitolo studia è più

generale: l'irreversibilità come struttura dei sistemi dinamici con biforcazioni, e la connessione tra irreversibilità e le impossibilità strutturali delle famiglie precedenti.

Definizione (irreversibilità strutturale). Un sistema dinamico  $D = (S, T)$  con spazio degli stati  $S$  e operatore di transizione  $T: S \rightarrow S$  è strutturalmente irreversibile se esiste uno stato  $s^* \in S$  — una biforcazione — tale che:

$$\exists s_1 \neq s_2 \in S : T^n(s_1) = T^n(s_2) = s^* \text{ per qualche } n \in \mathbb{N}$$

In altri termini, due stati distinti convergono allo stesso stato sotto l'evoluzione — l'operatore  $T$  non è iniettivo. Se  $T$  non è iniettivo,  $T^{-1}$  non è definita come funzione — la dinamica inversa non è deterministica, e il sistema non "ricorda" da quale dei due stati iniziali è venuto.

Questa è la struttura formale del ratchet: due stati distinti (la classe zero della popolazione e la classe  $k = 1$ ) convergono allo stesso stato (la classe  $k = 1$  dopo l'estinzione della classe zero) e l'inversione richiederebbe sapere da quale dei due stati si è partiti — informazione che è irrecuperabilmente persa.

### III. La path dependence come categoria topologica

La path dependence — la dipendenza dell'esito finale dalla traiettoria percorsa, analizzata nel capitolo V del ciclo biologico nel contesto dei periodi critici — ha una struttura formale precisa che questa sezione sviluppa in modo più sistematico.

Un sistema dinamico  $D = (S, T_t)$  con evoluzione continua ( $T_t$  è una famiglia di operatori parametrizzata dal tempo  $t \geq 0$ ) è path-dependent se esiste una funzione di "storia"  $H$  che mappa le traiettorie in insiemi di possibili stati futuri, e  $H$  non è costante sulle traiettorie con la stessa condizione iniziale.

Formalmente: siano  $\gamma_1, \gamma_2: [0, t_0] \rightarrow S$  due traiettorie con  $\gamma_1(0) = \gamma_2(0) = s_0$  (stessa condizione iniziale) ma  $\gamma_1 \neq \gamma_2$  (traiettorie diverse a causa di perturbazioni diverse durante  $[0, t_0]$ ). La path dependence è la proprietà:

$$\text{Reach}(\gamma_1, t > t_0) \neq \text{Reach}(\gamma_2, t > t_0)$$

dove  $\text{Reach}(\gamma, t)$  è l'insieme degli stati raggiungibili dalla traiettoria  $\gamma$  nel futuro dopo il tempo  $t$ . Le traiettorie diverse nel passato portano a insiemi di futuri diversi — la storia conta.

La struttura topologica della path dependence può essere catturata attraverso la nozione di bacino di attrazione (basin of attraction) di un attrattore. Sia  $A$  un attrattore di  $D$  — uno stato stazionario, un ciclo limite, o un attrattore strano. Il bacino di attrazione di  $A$  è:

$$B(A) = \{s \in S : T_t(s) \rightarrow A \text{ per } t \rightarrow \infty\}$$

Un sistema con più attrattori  $A_1, A_2, \dots$  ha la struttura dello spazio degli stati diviso in bacini di attrazione  $\{B(A_1), B(A_2), \dots\}$ . La path dependence è la struttura di questi bacini: quale bacino il sistema occupa dipende dalla traiettoria percorsa nelle fasi iniziali dell'evoluzione.

Per i sistemi biologici con periodi critici, la struttura dei bacini di attrazione ha una caratteristica specifica: i bacini cambiano forma nel tempo. Durante il periodo critico, i bacini sono ampi e le transizioni tra di essi sono facili (alta temperatura termodinamica, alta plasticità sinaptica). Dopo la chiusura del periodo critico, i bacini si restringono e le barriere tra essi crescono (bassa temperatura, bassa plasticità). La path dependence che emerge è la traiettoria del sistema nello spazio degli stati durante il periodo in cui le barriere erano ancora basse — quella traiettoria determina quale bacino il sistema occupa per il resto della sua vita, con barriere ora troppo alte per essere attraversate spontaneamente.

#### IV. La termodinamica dell'informazione e il principio di Landauer

La connessione tra irreversibilità e impossibilità strutturali passa attraverso il principio di Landauer (1961) — uno dei risultati più profondi nella termodinamica dell'informazione, che connette la computazione alla fisica.

Principio di Landauer. Qualsiasi operazione computazionale logicamente irreversibile — qualsiasi operazione che cancella informazione — deve dissipare nella sua realizzazione fisica almeno  $k_B T \ln 2$  di energia per bit cancellato, dove  $k_B$  è la costante di Boltzmann e  $T$  è la temperatura assoluta.

La dimostrazione del principio usa la corrispondenza tra informazione e entropia. La cancellazione di 1 bit riduce l'informazione del sistema di 1 bit — riduce l'entropia del sistema di  $k_B \ln 2$ . Per il secondo principio, la riduzione dell'entropia del sistema deve essere compensata da un aumento dell'entropia

dell'ambiente di almeno  $k_B \ln 2$  — il che richiede la dissipazione di almeno  $k_B T \ln 2$  di energia sotto forma di calore nell'ambiente.

Il principio di Landauer ha una formulazione operativa importante: le operazioni reversibili (che non cancellano informazione) possono essere realizzate in linea di principio senza dissipazione energetica; le operazioni irreversibili richiedono necessariamente dissipazione. La distinzione tra computazione reversibile e irreversibile è quindi una distinzione termodinamica — le operazioni irreversibili aumentano l'entropia dell'universo.

La connessione con le impossibilità strutturali è la seguente: tutte le impossibilità che abbiamo analizzato sono impossibilità di "cancellazione informatica". La diagonalizzazione mostra che certi fatti (le verità semantiche) non possono essere "compressi" nella struttura sintattica del sistema senza perdita irreversibile di informazione. Le impossibilità dimensionali mostrano che la proiezione di  $m$  feature in  $n < m$  dimensioni cancella irreversibilmente le componenti lungo le  $(m-n)$  dimensioni mancanti. Le impossibilità di complessità mostrano che la computazione di problemi NP-hard richiede risorse che, se risparmiate attraverso approssimazioni, producono perdita irreversibile di precisione nella soluzione.

In tutti questi casi, la "cancellazione" è ciò che produce l'impossibilità: non si può avere sia la rappresentazione completa sia la dimensione finita; non si può avere sia la semantica completa sia la sintassi finita; non si può avere sia la soluzione ottimale sia il tempo polinomiale. Il principio di Landauer formalizza perché queste impossibilità hanno un costo energetico reale: la loro violazione richiederebbe la cancellazione di informazione senza costo — un'operazione fisicamente impossibile.

## **V. Il ratchet di Müller come impossibilità di terzo tipo**

Il ratchet di Müller (Capitolo II del ciclo biologico) è una delle impossibilità biologiche più eleganti, e nell'ambito di questo ciclo assume un significato più profondo: è la manifestazione biologica dell'irreversibilità del secondo principio nella struttura delle popolazioni genetiche.

Ricordiamo la struttura formale: in una popolazione asessuata di dimensione  $N$  con tasso di mutazione  $\mu$ , la classe degli individui con il minor numero di

mutazioni deleterie (la "classe zero") si estingue per deriva genetica con probabilità positiva in ogni generazione. Una volta estinta, non può essere ricostituita — il ratchet è scattato. La velocità del ratchet è approssimativamente:

$$v \approx \mu \exp(-U/s) / T$$

dove  $U = N\mu$  è il numero atteso di mutazioni per genoma per generazione,  $s$  è l'effetto selettivo di ciascuna mutazione, e  $T$  è il tempo medio di estinzione della classe zero.

La struttura dell'irreversibilità del ratchet di Müller è quella del principio di Landauer applicato alla genetica:

L'informazione "genoma con zero mutazioni" viene cancellata quando la classe zero si estingue. Questa cancellazione è fisicamente irreversibile nel senso del principio di Landauer: richiederebbe che tutte le mutazioni nelle posizioni specifiche del genoma venissero reversate simultaneamente, il che richiederebbe un abbassamento impossibile dell'entropia genetica senza una compensazione entropica nel resto del sistema.

La connessione formale si stabilisce attraverso la nozione di entropia genetica. Sia  $p_k$  la frequenza degli individui con  $k$  mutazioni nella popolazione. L'entropia genetica è:

$$H_{\text{gen}} = -\sum_k p_k \log p_k$$

Quando il ratchet scatta, l'entropia genetica aumenta (la distribuzione delle mutazioni si sposta verso valori più alti) — esattamente come l'entropia termodinamica aumenta in un processo irreversibile. La velocità dell'aumento di entropia genetica per scatto del ratchet è:

$$\Delta H_{\text{gen}} = H(\{p_0=0, p_1, p_2, \dots\}) - H(\{p_0, p_1, p_2, \dots\}) > 0$$

L'irreversibilità del ratchet è la versione genetica del secondo principio — e il sesso (la soluzione evolutiva al ratchet) è la versione biologica della computazione reversibile: la ricombinazione sessuale permette di ricostruire la classe zero senza cancellazione irreversibile dell'informazione genetica, esattamente come la computazione reversibile permette di eseguire operazioni logiche senza dissipazione energetica.

## VI. I periodi critici come transizioni di fase irreversibili

La chiusura dei periodi critici (Capitolo V del ciclo biologico) è un'altra manifestazione della quinta famiglia, e la sua struttura formale rivela la connessione tra irreversibilità e la fisica delle transizioni di fase.

Il modello formale della chiusura del periodo critico era la dinamica di Langevin con temperatura decrescente:

$$dx/dt = -\nabla F(x) + \sqrt{2T(t)} \eta(t)$$

con  $T(t) \rightarrow 0$  per  $t \rightarrow \infty$ . Questo modello è precisamente quello del ricottura simulata (simulated annealing) — un algoritmo di ottimizzazione che abbassa gradualmente la temperatura per trovare il minimo globale di una funzione di costo  $F$ . La chiusura del periodo critico è il "congelamento" del sistema in un minimo di  $F$  — la cristallizzazione della struttura neurale in un bacino di attrazione specifico.

La connessione con le transizioni di fase della fisica statistica è esplicita. Una transizione di fase di secondo ordine (come la magnetizzazione di un ferromagnete al di sotto della temperatura di Curie) ha la struttura:

$T > T_c$ : una fase (paramagnetica, singolo minimo di  $F$ )  $T < T_c$ : due fasi (ferromagnetica, due minimi di  $F$ )

La transizione è reversibile sopra  $T_c$  (il sistema è nel paramagnetico unico) ma irreversibile sotto  $T_c$  (il sistema è magnetizzato in una direzione o nell'altra, e invertire richiede attraversare la barriera di energia tra i due minimi). Analogamente, il periodo critico è la fase "paramagnetica" dello sviluppo neurale — il sistema può essere modificato facilmente in qualsiasi direzione. La chiusura del periodo critico è la transizione sotto  $T_c$  — il sistema si "magnetizza" in una particolare configurazione neurale e le barriere che separano quella configurazione dalle alternative diventano insorpassabili.

Teorema (irreversibilità della chiusura del periodo critico). Sia il sistema neurale governato dalla dinamica di Langevin con  $T(t) \rightarrow 0$ . Per qualsiasi programma di raffreddamento  $T(t)$  strettamente decrescente con  $T(t) \rightarrow 0$  esponenzialmente, il sistema raggiunge asintoticamente un minimo locale di  $F$  con probabilità  $1 - O(\exp(-\Delta F_{\min}/T_{\text{final}}))$ .

Questo teorema (dimostrato in forma analoga nel capitolo V del ciclo biologico) ha una struttura identica al teorema di impossibilità dell'estrazione del lavoro da una fonte termica fredda: nessuna macchina termica può estrarre lavoro da una riserva a temperatura  $T_{\text{final}}$  senza aumentare l'entropia dell'universo. Analogamente, nessun processo di plasticità sinaptica può "estrarre" una riconfigurazione neurale da un sistema congelato a  $T \approx 0$  senza un input energetico che aumenta l'entropia del sistema (che si traduce concretamente nell'indebolimento delle strutture già consolidate per costruirne di nuove).

## VII. La deriva epigenetica e l'entropia dell'informazione regolatoria

La deriva epigenetica — l'accumulo progressivo di errori nel codice epigenetico durante le divisioni cellulari (analizzata nel capitolo XVI del ciclo biologico) — è un'altra istanza della quinta famiglia, con una struttura particolarmente chiara in termini di informazione teorica.

Il codice epigenetico di una cellula è una struttura di informazione che specifica quali geni sono attivi e quali sono silenziati — una struttura che è trasmessa attraverso le divisioni cellulari con fedeltà  $p_{\text{err}}$  per sito per divisione. Dopo  $n$  divisioni, la fedeltà complessiva del sito  $i$  è  $(1-p_{\text{err}})^n$ . Per  $p_{\text{err}} = 0.02$  e  $n = 50$  divisioni:

$$P_{\text{corretto}}(50) = (0.98)^{50} \approx 0.36$$

L'entropia del codice epigenetico per sito dopo  $n$  divisioni è:

$$H_{\text{epigenetic}}(n) = -P_{\text{corretto}}(n) \log P_{\text{corretto}}(n) - P_{\text{errore}}(n) \log P_{\text{errore}}(n)$$

con  $P_{\text{errore}}(n) = 1 - P_{\text{corretto}}(n) = 1 - (1-p_{\text{err}})^n$ . Per  $n = 50$ :

$$H_{\text{epigenetic}}(50) = -0.36 \log(0.36) - 0.64 \log(0.64) \approx 0.95 \text{ bit}$$

L'entropia del sito epigenetico è quasi massima (1 bit) dopo 50 divisioni — l'informazione sulla metilazione originale è quasi completamente perduta. L'irreversibilità della deriva epigenetica è la versione cellulare del principio di Landauer: ogni divisione cellulare "cancella" parzialmente l'informazione

epigenetica, e questa cancellazione aumenta l'entropia del sistema in modo irreversibile.

La struttura della deriva epigenetica come impossibilità strutturale è la seguente: la cellula non può "sapere" se il suo codice epigenetico corrente è quello originale o una versione degradata, perché l'informazione sul codice originale è stata parzialmente cancellata dalla deriva. Questa impossibilità è analoga all'impossibilità di verificare la propria semantica dall'interno (Capitolo III): la cellula non può accedere al "significato" del suo codice epigenetico (il tipo cellulare per cui è programmata) attraverso la sola struttura del codice corrente (la sintassi epigenetica attuale).

### **VIII. La struttura categoriale dell'irreversibilità: i morfismi che non ammettono inversi**

La quinta famiglia può essere formalizzata categorialmente in modo da connetterla alle altre famiglie attraverso il linguaggio della teoria delle categorie già usato per il teorema di Lawvere.

In una categoria  $C$ , un morfismo  $f: A \rightarrow B$  è un isomorfismo se esiste un morfismo inverso  $g: B \rightarrow A$  tale che  $g \circ f = \text{id}_A$  e  $f \circ g = \text{id}_B$ . Un morfismo che non è un isomorfismo è irreversibile — la trasformazione da  $A$  a  $B$  non può essere disfatta.

L'irreversibilità degli isomorfismi in categorie biologiche ha una struttura precisa. Nella categoria  $\text{Set}$  (insiemi e funzioni), tutte le funzioni iniettive (one-to-one) sono reversibili (ammettono una funzione inversa sinistra); le funzioni non-iniettive (che "collassano" elementi distinti) non lo sono. Il ratchet di Müller è un morfismo non-iniettivo nella categoria degli stati genetici: due stati distinti (la classe zero e la classe  $k=1$ ) vengono mappati allo stesso stato (la classe  $k=1$  dopo l'estinzione della classe zero). La non-iniettività produce l'irreversibilità.

Nella categoria degli spazi di Hilbert (spazi di funzioni, usati nella meccanica quantistica), l'evoluzione temporale è unitaria — un isomorfismo, quindi reversibile. La misurazione quantistica, al contrario, è un morfismo non-unitario che proietta lo stato quantistico su un sottospazio — non è invertibile. La dissipazione quantistica (l'interazione con l'ambiente, il "decoherence") è

l'equivalente quantistico del ratchet di Müller: produce morfismi non-unitari che collassano stati distinti sullo stesso stato risultante.

Il quadro categoriale unifica quindi l'irreversibilità biologica, termodinamica e quantistica come la stessa struttura astratta: morfismi non-iniettivi (o non-unitari) in categorie appropriate. La quinta famiglia delle impossibilità strutturali è la famiglia dei morfismi che non ammettono inversi — la famiglia dei processi che "dimenticano" la propria origine.

## IX. Il teorema dell'entropia cognitiva

La connessione tra la quinta famiglia e le altre quattro si manifesta in quello che propongo di chiamare il teorema dell'entropia cognitiva — un risultato che unifica le impossibilità di irreversibilità, dimensionalità e di chiusura semantica in una relazione formale unica.

Definizione (entropia cognitiva). Sia  $S$  un sistema cognitivo con modello generativo  $q$  del mondo. L'entropia cognitiva di  $S$  è:

$$H_{\text{cog}}(S) = H(q) - I(q; M)$$

dove  $H(q)$  è l'entropia del modello generativo (la complessità del modello),  $I(q; M)$  è l'informazione mutua tra il modello  $q$  e il mondo reale  $M$  (la quantità di informazione sul mondo che il modello cattura), e  $H_{\text{cog}}$  è il gap informativo — la differenza tra la complessità del modello e la sua accuratezza nel rappresentare il mondo.

Questa definizione è formalmente identica al gap informativo del capitolo III:  $H_{\text{cog}}(S) = \text{Gap}(\Sigma_q, M)$  dove  $\Sigma_q$  è la "sintassi" del modello e  $M$  è la "semantica" del mondo. La novità è la dimensione temporale:  $H_{\text{cog}}$  cambia nel tempo in risposta all'esperienza e all'apprendimento.

Teorema (monotonia dell'entropia cognitiva). Sia  $S$  un sistema cognitivo con dinamica di aggiornamento governata da qualsiasi processo di apprendimento bayesiano (minimizzazione dell'energia libera variazionale, gradient descent, TD learning, o qualsiasi altro). Allora:

$$E[H_{\text{cog}}(S, t+1)] \leq H_{\text{cog}}(S, t) + \varepsilon(t)$$

dove  $\epsilon(t) \geq 0$  è il termine di "entropia creativa" — l'aumento di entropia cognitiva dovuto ai cambiamenti del mondo  $M$  che il sistema non può anticipare.

La disuguaglianza afferma che l'entropia cognitiva non aumenta in attesa in assenza di cambiamenti imprevisi del mondo — l'apprendimento riduce  $H_{\text{cog}}$ . Ma  $\epsilon(t) \geq 0$  garantisce che i cambiamenti imprevisi del mondo possono sempre aumentare  $H_{\text{cog}}$  — non è possibile mantenere  $H_{\text{cog}} = 0$  in un mondo che cambia.

Dimostrazione (sketch). Il processo di apprendimento bayesiano minimizza l'energia libera variazionale  $F = H_{\text{cog}} + \text{const}$  (per la decomposizione  $F = \text{Complessità} - \text{Accuratezza}$  del capitolo VIII del ciclo biologico). La minimizzazione di  $F$  in risposta all'evidenza corrente non può aumentare  $H_{\text{cog}}$  condizionatamente alla struttura corrente del mondo  $M$ . Ma se  $M$  cambia in modo impreveduto (da  $M$  a  $M'$  con  $D_{\text{KL}}(M \parallel M') > 0$ ), l'aggiornamento del modello non può anticipare il cambiamento, e l'entropia cognitiva aumenta della quantità  $\epsilon(t) = D_{\text{KL}}(M' \parallel q)$  — la distanza KL tra il nuovo stato del mondo e il modello corrente.  $\square$

Il teorema dell'entropia cognitiva è la versione cognitiva del secondo principio della termodinamica: come l'entropia termodinamica non decresce in un sistema isolato, l'entropia cognitiva non decresce in un agente che non apprende; e come il secondo principio richiede input energetico per ridurre l'entropia locale, il teorema richiede apprendimento (input informativo) per ridurre  $H_{\text{cog}}$ .

Ma il teorema ha anche una componente di irreversibilità:  $\epsilon(t) \geq 0$  implica che qualsiasi cambiamento impreveduto del mondo aumenta  $H_{\text{cog}}$  in modo che non può essere completamente compensato dall'apprendimento successivo senza passare per una fase di incertezza aumentata. Questo è il meccanismo dell'obsolescenza cognitiva — l'inevitabile gap tra il modello del mondo del sistema cognitivo e il mondo reale in rapida evoluzione.

## **X. L'irreversibilità come condizione della novità**

Le quattro famiglie di impossibilità precedenti erano, in un certo senso, chiuse: definivano confini fissi che nessun sistema può attraversare. La quinta famiglia — l'irreversibilità — introduce una dimensione temporale che le altre non hanno: le

impossibilità irreversibili definiscono confini che cambiano nel tempo e che il sistema percorre in una sola direzione.

Questa asimmetria temporale è la condizione della novità. La storia esiste soltanto perché i processi sono irreversibili — se ogni processo fosse invertibile, non vi sarebbe freccia del tempo, non vi sarebbe passato distinto dal futuro, non vi sarebbero eventi irreversibili come la nascita, la morte, l'apprendimento, il dimenticare. La novità — la comparsa di strutture che non esistevano prima — è possibile soltanto in un universo con irreversibilità termodinamica.

Il ratchet di Müller è il prezzo evolutivo di questa novità: l'accumulo delle mutazioni è irreversibile, il che significa che ogni generazione è genuinamente diversa dalle precedenti — non soltanto in modo casuale, ma in modo che non può essere ripercorso a ritroso. Allo stesso modo, la chiusura del periodo critico è il prezzo cognitivo della novità: la plasticità si chiude affinché le strutture apprese siano stabili, e questa stabilità è la condizione sotto cui l'identità cognitiva diventa possibile. Un sistema interamente plastico — senza irreversibilità — non può avere identità nel tempo.

La quinta famiglia di impossibilità è quindi quella che connette le impossibilità strutturali alla struttura del tempo, dell'identità e della storia. Le impossibilità diagonali e di chiusura semantica definiscono i confini dello spazio dei problemi rappresentabili; le impossibilità dimensionali definiscono i confini della fedeltà della rappresentazione; le impossibilità di complessità definiscono i confini della velocità di raggiungimento delle soluzioni; le impossibilità di irreversibilità definiscono i confini della direzione del tempo nella cognizione e nella vita. Insieme, le cinque famiglie delineano la topologia completa del territorio in cui il pensiero è possibile — e del confine oltre il quale non lo è più.

---

## Capitolo VII

### L'unità delle cinque famiglie

#### La derivazione delle impossibilità strutturali come corollari del confine tarskiano e la struttura di un principio unificante

##### Mappa epistemica — Capitolo VII

| Status     | Cod. | Claim                                                                                                                         |
|------------|------|-------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO | D    | Le condizioni C1 e C2 del PAI seguono formalmente dal teorema di Tarski sotto le ipotesi H1 e H2.                             |
| DIMOSTRATO | D    | Le implicazioni $F_1 \rightarrow F_2$ (via Brouwer) e $F_4 \rightarrow F_5$ (via Landauer) sono dimostrate.                   |
| CONGETTURA | C    | Il PAI nella forma generale a quattro condizioni separate: dipende da $P \neq NP$ . Stesso status epistemico di $P \neq NP$ . |
| CONGETTURA | C    | La legge di conservazione della carica impossibilitante: congettura strutturale. Solo due istanze parziali sono dimostrate.   |

### I. Il problema dell'unità

I sei capitoli precedenti hanno costruito una tassonomia articolata — cinque famiglie di impossibilità strutturali con strumenti formali distinti, orizzonti applicativi diversi, e storie intellettuali separate. Il catalogo è ricco e, spero, illuminante. Ma un catalogo non è una teoria. Una teoria richiede più di un'enumerazione di casi: richiede un principio da cui i casi si derivano, una struttura che li generi invece di semplicemente raccogliarli.

il capitolo III ha già avanzato la tesi centrale: il confine sintassi-semantica è il locus generativo di tutte le impossibilità strutturali. i capitoli successive hanno suffragato questa tesi con analisi sistematiche per ciascuna famiglia, ma la dimostrazione formale dell'unità — la derivazione delle famiglie come corollari di un unico principio — è rimasta promessa. questo capitolo la mantiene.

L'argomento procede in tre movimenti. Il primo identifica il principio unificante nella forma più astratta possibile — una proposizione sulle strutture auto-rappresentazionali che implica, come corollari, le impossibilità di tutte e

cinque le famiglie. Il secondo movimento dimostra i corollari con la precisione che il programma del ciclo richiede. Il terzo movimento affronta il problema epistemologico inverso: se le cinque famiglie sono corollari di un unico principio, perché sono apparse storicamente come risultati separati in discipline separate? La risposta a questa domanda rivela qualcosa di non ovvio sulla struttura della conoscenza scientifica stessa.

## II. Il principio di auto-rappresentazione insufficiente

### [CONGETTURA FONDATA]

Chiarimento epistemico fondamentale. Il PAI non è un teorema matematico nel senso dei teoremi di Gödel, Tarski e Turing. È quello che potremmo chiamare un 'teorema filosofico con dimostrazione condizionale': la sua parte formale (le condizioni C1 e C2) segue dal teorema di Tarski sotto H1 e H2; la separazione tra C3 (risorse superpolinomiali) e C4 (perdita di informazione) dipende da  $P \neq NP$ .

Se  $P = NP$ , l'argomento per assurdo non distinguerebbe C3 da C4, e il PAI nella sua forma a quattro condizioni separate non sarebbe formalmente derivabile. Il PAI viene presentato come principio unificante potente e quasi certamente vero — con lo stesso grado di certezza di  $P \neq NP$ .

Il principio unificante che propongo può essere enunciato in forma preliminare come segue:

Principio di auto-rappresentazione insufficiente (PAI). Sia  $S$  un sistema dotato di una struttura  $S_{int}$  di rappresentazione interna (la "sintassi" del sistema) e di una struttura  $S_{ext}$  del dominio che il sistema descrive o in cui opera (la "semantica" del sistema). Se  $S_{int}$  è sufficientemente ricca da esprimere asserzioni non-banali su  $S_{ext}$ , e se  $S$  tenta di catturare  $S_{ext}$  completamente attraverso  $S_{int}$ , allora  $S$  incontra necessariamente almeno una delle seguenti condizioni: (I) genera contraddizioni; (II) è incompleta rispetto a  $S_{ext}$ ; (III) richiede risorse che eccedono quelle disponibili a qualsiasi sistema fisicamente realizzabile; (IV) produce perdite di informazione irreversibili.

Questa enunciazione è ancora troppo informale per servire come principio formale: i termini "sufficientemente ricca", "asserzioni non-banali", "catturare completamente" richiedono precisazione. Il paragrafo seguente costruisce la versione rigorosa.

Definizione (sistema auto-rappresentazionale). Un sistema auto-rappresentazionale è una coppia  $(S, \varphi)$  dove  $S$  è un sistema formale (un linguaggio con assiomi e

regole di inferenza) e  $\varphi: S \rightarrow S$  è una funzione auto-referenziale — una funzione che mappa strutture di  $S$  in strutture di  $S$ , in modo tale che  $S$  possa esprimere asserzioni sulle strutture prodotte da  $\varphi$ . La funzione  $\varphi$  è l'operazione di "auto-applicazione" del sistema.

Esempi di sistemi auto-rappresentazionali: (PA, diag) dove diag è la funzione di diagonalizzazione gödeliana;  $(\mathbb{R}^n, \text{Proj})$  dove Proj è la proiezione da  $\mathbb{R}^m$  a  $\mathbb{R}^n$  con  $m > n$ ; (Programmi, Eval) dove Eval è l'interprete universale che valuta programmi su input; (GRN, Reg) dove Reg è la rete di regolazione che trasforma stati cellulari in altri stati cellulari; (Modello\_generativo, Min\_F) dove Min\_F è l'operazione di minimizzazione dell'energia libera.

Definizione (rappresentazione completa). Una funzione  $\varphi: S \rightarrow S$  realizza una rappresentazione completa di  $S_{\text{ext}}$  se per ogni struttura  $s_{\text{ext}} \in S_{\text{ext}}$  esiste un elemento  $s \in S$  tale che  $\varphi(s)$  corrisponde a  $s_{\text{ext}}$  — ossia  $\varphi$  è suriettiva rispetto a  $S_{\text{ext}}$  nella topologia rilevante per il problema considerato.

Il principio di auto-rappresentazione insufficiente nella sua forma rigorosa è allora:

Teorema (PAI, forma rigorosa). Sia  $(S, \varphi)$  un sistema auto-rappresentazionale con  $\varphi: S \rightarrow S$ . Se  $\varphi$  è suriettiva rispetto a  $S_{\text{ext}}$  (realizza una rappresentazione completa) e se  $S$  contiene abbastanza struttura da esprimere asserzioni non-banali su se stesso (contiene l'equivalente della codifica gödeliana), allora almeno una delle seguenti condizioni vale:

- (I)  $S$  è inconsistente:  $S \vdash \perp$  (il sistema deriva la contraddizione)
- (II)  $S$  è incompleto: esiste una proposizione  $\varphi \in \text{Prop}(S_{\text{ext}})$  tale che  $S \not\vdash \varphi$  e  $S \not\vdash \neg\varphi$
- (III) La computazione di  $\varphi$  richiede risorse superpolinomiali rispetto alla dimensione dell'input:  $\varphi \notin \text{PTIME}$
- (IV) La composizione iterata  $\varphi^n$  con  $n \rightarrow \infty$  produce perdita di informazione:  $H(\varphi^n(s)) < H(\varphi(s))$  per qualche  $s$ ,  $n$  sufficientemente grandi

Dimostrazione. La dimostrazione procede per assurdo. Si supponga che nessuna delle quattro condizioni valga:  $S$  è consistente, completo,  $\varphi$  è calcolabile in tempo polinomiale, e  $\varphi$  preserva l'informazione esattamente in ogni iterazione. Sotto queste ipotesi, si costruisce la seguente struttura:

Poiché  $\varphi$  è suriettiva e calcolabile in tempo polinomiale, esiste un algoritmo  $A$  che, dato qualsiasi elemento di  $S_{\text{ext}}$ , trova il corrispondente elemento di  $S$  in tempo polinomiale — un algoritmo di "inversione" di  $\varphi$ .

Poiché  $S$  è completo, ogni proposizione su  $S_{\text{ext}}$  è decidibile in  $S$ .

Poiché  $\varphi$  preserva l'informazione, la composizione  $\varphi^{-1} \circ \varphi = \text{id}$  — ogni elemento di  $S$  ha un unico corrispondente in  $S_{\text{ext}}$  che è recuperabile.

Dall'esistenza di  $A$  (inversione polinomiale di  $\varphi$ ) e dalla completezza di  $S$ , si costruisce il seguente programma che calcola la funzione di verità di  $S$ :

Per ogni proposizione  $p$  di  $S$ , calcola  $\varphi^{-1}(p)$  (polinomiale per ipotesi), poi usa la completezza per decidere la verità di  $\varphi^{-1}(p)$  in  $S_{\text{ext}}$ .

Ma questo è un predicato di verità per  $S$  definito internamente — contraddicendo il teorema di Tarski, che è applicabile perché  $S$  contiene l'equivalente della codifica gödeliana per ipotesi. La contraddizione mostra che le quattro condizioni non possono essere simultaneamente false.  $\square$

Il PAI dimostra che le cinque famiglie non sono indipendenti — sono le quattro uscite possibili dalla medesima struttura impossibile. Ogni sistema auto-rappresentazionale che cerca la completezza rispetto al proprio dominio finisce in una delle quattro condizioni (I)-(IV), corrispondenti esattamente alle quattro famiglie originali (più la quinta, l'irreversibilità, che si identifica con la perdita di informazione della condizione IV).

### III. La derivazione della Famiglia I come corollario del PAI

La Famiglia I — le impossibilità diagonali — è il corollario più diretto del PAI. La sua derivazione recupera la struttura del teorema di Lawvere (Capitolo II) ma in modo integrato con la struttura generale.

Corollario I (impossibilità diagonale). Sotto le ipotesi del PAI, se  $S$  è consistente (condizione I non verificata) e  $\varphi$  è calcolabile in tempo polinomiale (condizione III non verificata), allora  $S$  è necessariamente incompleto (condizione II).

Derivazione. Per le ipotesi del corollario,  $S$  è consistente e  $\varphi$  è trattabile. Il PAI garantisce che almeno una delle quattro condizioni vale; le ipotesi escludono I e

III. Rimangono II e IV. Se IV non vale ( $\varphi$  preserva l'informazione), allora II deve valere:  $S$  è incompleto.

La costruzione esplicita della proposizione indecidibile segue il lemma della diagonalizzazione: poiché  $S$  contiene la codifica gödeliana, esiste una sentenza  $G$  tale che  $S \vdash G \leftrightarrow \neg \text{Provabile}_S(\ulcorner G \urcorner)$ . Se  $S$  è consistente,  $G$  è indimostrabile; se  $S$  è  $\omega$ -consistente,  $\neg G$  è indimostrabile.  $\square$

La struttura di questo corollario è: la consistenza e la trattabilità computazionale forzano l'incompletezza. Un sistema che è sia consistente sia trattabile computazionalmente non può essere completo. Questo collega la Famiglia I alla Famiglia III in modo formale: la completezza richiede o l'abbandono della consistenza (generare contraddizioni) o l'abbandono della trattabilità computazionale (richiedere risorse superpolinomiali).

La manifestazione biologica di questo corollario è nel divario tra i sistemi cognitivi biologici (che sono approssimativamente consistenti e approssimativamente trattabili computazionalmente) e la loro necessaria incompletezza rispetto al mondo: nessun sistema cognitivo biologico ha un modello completo del proprio ambiente. L'incompletezza non è un difetto evolutivo — è la conseguenza necessaria della consistenza e della trattabilità.

#### **IV. La derivazione della Famiglia II come corollario del PAI**

Corollario II (impossibilità dimensionale). Sotto le ipotesi del PAI, se  $S$  è consistente, completo e trattabile computazionalmente (condizioni I, II, III non verificate), allora  $\varphi$  non può essere sia suriettiva sia isometrica — qualsiasi rappresentazione completa introduce distorsione geometrica (condizione IV nella sua forma metrica).

Derivazione. Per le ipotesi del corollario,  $S$  è consistente (non I), completo (non II), e  $\varphi$  è trattabile (non III). Il PAI garantisce che la condizione IV deve valere:  $\varphi$  produce perdita di informazione. Nel caso in cui  $S_{\text{int}}$  è uno spazio metrico (ad esempio,  $\mathbb{R}^n$ ) e  $S_{\text{ext}}$  è uno spazio metrico di dimensione superiore ( $\mathbb{R}^m$  con  $m > n$ ), la perdita di informazione della condizione IV si manifesta come distorsione delle distanze — l'impossibilità di proiettare  $\mathbb{R}^m$  in  $\mathbb{R}^n$  preservando tutte le distanze.

Formalmente: se  $\varphi: \mathbb{R}^m \rightarrow \mathbb{R}^n$  è la proiezione di rappresentazione e  $m > n$ , allora per qualsiasi  $\varphi$  suriettiva:

$$\exists x, y \in \mathbb{R}^m : \|\varphi(x) - \varphi(y)\| \neq \|x - y\|$$

La dimostrazione usa la conservazione della dimensione di Hausdorff sotto isometrie: se  $\varphi$  fosse isometrica,  $\dim_H(\varphi(\mathbb{R}^m)) = \dim_H(\mathbb{R}^m) = m$ . Ma  $\dim_H(\varphi(\mathbb{R}^m)) \leq \dim_H(\mathbb{R}^n) = n < m$  — contraddizione.  $\square$

Il corollario II mostra che la Famiglia II è la Famiglia IV (perdita di informazione) specializzata al caso in cui lo spazio interno  $S_{\text{int}}$  è uno spazio metrico. La distorsione dimensionale è la forma che la perdita di informazione assume quando il sistema di rappresentazione è uno spazio vettoriale euclideo. L'incompletezza della Famiglia I è la forma che la perdita di informazione assume quando il sistema è un sistema logico formale.

## V. La derivazione della Famiglia III come corollario del PAI

Corollario III (impossibilità di complessità). Sotto le ipotesi del PAI, se  $S$  è consistente e completo (condizioni I, II non verificate), e se  $\varphi$  non produce perdita di informazione irreversibile (condizione IV non verificata), allora  $\varphi$  non è calcolabile in tempo polinomiale (condizione III).

Derivazione. Per le ipotesi,  $S$  è consistente (non I), completo (non II), e  $\varphi$  è isometrica rispetto all'informazione (non IV). Il PAI garantisce che la condizione III deve valere:  $\varphi$  richiede risorse superpolinomiali.

La costruzione esplicita usa il fatto che la completezza con conservazione dell'informazione implica la capacità di rispondere a qualsiasi domanda su  $S_{\text{ext}}$  — incluse le domande NP-complete. Se  $\varphi$  fosse calcolabile in tempo polinomiale e fosse completo (ogni domanda su  $S_{\text{ext}}$  è decidibile), allora il decisore potrebbe:

Trasformare qualsiasi istanza di 3-SAT in un elemento di  $S$  (usando  $\varphi$  in tempo polinomiale)

Decidere la sua soddisfacibilità in  $S$  (usando la completezza)

Questo darebbe un algoritmo polinomiale per 3-SAT, contraddicendo la NP-completezza di 3-SAT sotto  $P \neq NP$ .  $\square$

Il corollario III mostra che la Famiglia III (impossibilità di complessità) è la conseguenza di richiedere simultaneamente consistenza, completezza e conservazione dell'informazione: le tre condizioni insieme forzano l'intrattabilità computazionale. La congettura  $P \neq NP$  è quindi non soltanto una congettura sulla difficoltà degli algoritmi — è una conseguenza della struttura del PAI che afferma la non-vacuità delle impossibilità di complessità.

Questo produce una nuova prospettiva sulla congettura:  $P \neq NP$  è vera non perché i nostri algoritmi siano ingenui e potrebbero essere migliorati, ma perché la sua negazione ( $P = NP$ ) contraddirebbe il PAI — contraddirebbe la struttura auto-rappresentazionale fondamentale di qualsiasi sistema sufficientemente ricco. La dimostrazione di  $P \neq NP$ , se venisse trovata, dimostrerebbe in realtà il PAI applicato alle classi di complessità.

## VI. La derivazione della Famiglia IV come corollario del PAI

La Famiglia IV — le impossibilità di chiusura semantica — è la famiglia più direttamente connessa al PAI: il principio è formulato esattamente nei termini del confine sintassi-semantica che caratterizza questa famiglia. La derivazione è quindi quasi immediata.

Corollario IV (impossibilità di chiusura semantica). Sia  $(S, \varphi)$  un sistema auto-rappresentazionale dove  $S_{\text{ext}}$  contiene la semantica del sistema (i valori di verità delle proposizioni di  $S$  nei modelli di  $S_{\text{ext}}$ ). Se  $S$  è consistente e  $\varphi$  è la funzione che tenta di rappresentare la relazione di soddisfazione  $\models$  internamente a  $S$ , allora  $\varphi$  non è né totale (definita per tutte le proposizioni di  $S$ ) né coerente con la relazione  $\models$ .

Derivazione. Questa è esattamente la dimostrazione del teorema di Tarski (Capitolo III). La funzione  $\varphi$  che tenta di definire  $\text{Sat}(x, y)$  (la soddisfazione interna) produce, tramite il lemma della diagonalizzazione, una sentenza  $G$  tale che  $G \leftrightarrow \neg \text{Sat}(\ulcorner G \urcorner, \ulcorner M \urcorner)$ . La coerenza richiede che  $G$  sia vera nel modello se e soltanto se  $\text{Sat}(\ulcorner G \urcorner, \ulcorner M \urcorner)$  è vera nel modello; ma  $G$  afferma l'opposto di  $\text{Sat}(\ulcorner G \urcorner, \ulcorner M \urcorner)$ , generando contraddizione. Quindi  $\varphi$  non può essere la funzione di soddisfazione interna di  $S$ .  $\square$

La famiglia IV è la forma "pura" del PAI — il caso in cui la completezza auto-rappresentazionale è impossibile non perché si richiedono risorse eccessive o si produce distorsione metrica, ma perché il sistema stesso contraddice la propria semantica nell'atto di tentare di catturarla. È la famiglia in cui l'impossibilità ha il suo aspetto più essenziale.

## VII. La derivazione della Famiglia V come corollario del PAI

La quinta famiglia — l'irreversibilità — richiede la versione dinamica del PAI, in cui il sistema evolve nel tempo e la "completezza" è riferita alla rappresentazione dell'intero percorso storico, non soltanto dello stato corrente.

**Definizione** (sistema auto-rappresentazionale dinamico). Un sistema auto-rappresentazionale dinamico è una tripla  $(S, \varphi, T)$  dove  $(S, \varphi)$  è un sistema auto-rappresentazionale e  $T: S \rightarrow S$  è un operatore di evoluzione temporale che trasforma stati di  $S$  in stati successivi.

**Corollario V** (impossibilità di irreversibilità). Sia  $(S, \varphi, T)$  un sistema auto-rappresentazionale dinamico. Se  $T$  è non-iniettivo (due stati distinti possono convergere allo stesso stato) e se  $S$  tenta di rappresentare completamente la propria storia attraverso  $\varphi$  (la funzione di auto-rappresentazione dell'intera traiettoria), allora  $\varphi$  non può essere simultaneamente:

(a) suriettiva rispetto allo spazio delle traiettorie passate (b) Computabile in tempo finito a partire dallo stato corrente (c) Preservatrice dell'informazione storica completa

**Derivazione.** La non-iniettività di  $T$  implica che esiste una coppia di stati  $s_1 \neq s_2$  tali che  $T(s_1) = T(s_2) = s^*$ . A partire da  $s^*$ , non è possibile determinare se il sistema proveniva da  $s_1$  o da  $s_2$  soltanto a partire dallo stato corrente — l'informazione sulla differenza tra  $s_1$  e  $s_2$  è irrecuperabilmente perduta nella transizione  $T$ .

Se  $\varphi$  tentasse di recuperare questa informazione (condizione a — suriettiva sulle traiettorie), dovrebbe accedere a informazioni non presenti nello stato corrente  $s^*$ . Qualsiasi schema di recupero richiederebbe o:

- Risorse computazionali superpolinomiali nella dimensione dello spazio degli stati (violando b), perché il numero di possibili traiettorie che conducono a  $s^*$  cresce esponenzialmente con il tempo; oppure

- Una "memoria esterna" nello stato corrente  $s^*$  che codifichi l'intera storia — ma questa memoria avrebbe la stessa dimensione dello spazio delle traiettorie, il che non è compatibile con la rappresentazione all'interno di  $S$  di dimensione finita (violando  $c$  per il principio di Landauer: la codifica della storia richiederebbe cancellare informazione dal sistema per fare spazio alla memoria storica, aumentando irreversibilmente l'entropia).  $\square$

Il corollario V mostra che l'irreversibilità è il corollario del PAI nella sua dimensione temporale: un sistema che si evolve attraverso transizioni non-iniettive incontra necessariamente impossibilità di rappresentazione della propria storia che corrispondono alle condizioni (b) e (c) del PAI generale.

### VIII. La struttura dei corollari: un sistema di implicazioni

I cinque corollari non sono semplicemente derivati separatamente dal PAI — hanno struttura di sistema con implicazioni reciproche che rivelano la topologia interna delle famiglie di impossibilità.

#### *Proposizione (implicazioni tra famiglie). Sotto le ipotesi del PAI:*

(I  $\rightarrow$  II): Se un sistema evita le impossibilità diagonali (non produce contraddizioni attraverso auto-riferimento), incontra necessariamente le impossibilità di chiusura semantica o quelle di complessità.

(II  $\rightarrow$  III): Se un sistema evita le impossibilità di chiusura semantica (può rappresentare la propria semantica dall'interno), incontra necessariamente impossibilità di complessità.

(III  $\rightarrow$  IV): Se un sistema evita le impossibilità di complessità (può calcolare la propria semantica in tempo polinomiale), incontra necessariamente impossibilità di chiusura semantica o distorsioni dimensionali.

(IV  $\rightarrow$  V): Se un sistema evita le distorsioni dimensionali (mantiene rappresentazioni isometriche), incontra necessariamente perdite di informazione irreversibili nella sua evoluzione temporale.

Dimostrazione. Ogni implicazione è una forma del PAI con le ipotesi appropriate. La dimostrazione procede costruendo, per ciascuna implicazione, la struttura che forza l'impossibilità successiva quando la precedente è evitata.

Per  $(I \rightarrow II)$ : se il sistema è consistente (evita le contraddizioni della Famiglia I) e ha la proprietà di auto-riferimento necessaria per il PAI, allora per il corollario IV il sistema non può definire il proprio predicato di verità — l'impossibilità di chiusura semantica (Famiglia IV) è necessaria.

Per  $(II \rightarrow III)$ : se il sistema può rappresentare la propria semantica internamente (evita la Famiglia IV), quella rappresentazione deve essere calcolabile. Ma per il corollario III, una rappresentazione semantica completa calcolabile in tempo polinomiale contraddirebbe  $P \neq NP$ . Quindi la computazione della semantica deve richiedere risorse superpolinomiali — la Famiglia III è necessaria.

Per  $(III \rightarrow IV)$ : se il sistema calcola la propria semantica in tempo superpolinomiale (affronta la Famiglia III) ma non produce distorsioni dimensionali (evita la Famiglia II), deve preservare l'informazione isometricamente. La quantità di informazione preservata è dell'ordine di  $m$  bit per  $m$  feature, e la sua trasmissione nel tempo richiede un canale di capacità  $m$  bit. Ma se il sistema ha capacità di trasmissione  $n < m$  (dimensione fisica finita), la preservazione isometrica è impossibile — la Famiglia II emerge.

Per  $(IV \rightarrow V)$ : se il sistema ha dimensione fisica finita (affronta la Famiglia II) e la sua struttura interna rappresenta una quantità di informazione limitata, la sua evoluzione temporale deve essere non-iniettiva in qualche regime — stati diversi convergono allo stesso stato per perdita dell'informazione non rappresentata. La non-iniettività è la struttura della Famiglia V.  $\square$

La struttura di implicazioni  $I \rightarrow II \rightarrow III \rightarrow IV \rightarrow V$  non è circolare: è una catena che mostra come l'evitare una famiglia forza necessariamente l'incontro con la successiva. I sistemi biologici reali abitano questa catena in modo caratteristico: il cervello evita le contraddizioni interne (è approssimativamente consistente), incontra limiti di chiusura semantica (il divario esplicativo, il choice blindness), affronta impossibilità di complessità (usa euristiche invece dell'ottimizzazione esatta), ha distorsioni dimensionali (il binding problem, la sparsità), e subisce irreversibilità (la plasticità dipendente dall'esperienza, il dimenticare).

## IX. Il principio come legge di conservazione dell'impossibilità

Una struttura emerge dall'analisi dei corollari che merita di essere formulata esplicitamente: le impossibilità si conservano. Evitarne una equivale a incontrarne un'altra; minimizzarne una equivale ad amplificarne un'altra. Questo suggerisce l'esistenza di una quantità conservata — un'invariante delle impossibilità strutturali — che ogni sistema deve rispettare indipendentemente dalla propria architettura.

Propongo di chiamare questa quantità la carica impossibilitante di un sistema, denotata  $\Psi(S)$ . Informalmente,  $\Psi(S)$  misura la "quantità totale di impossibilità" che il sistema  $S$  deve incontrare data la sua struttura auto-rappresentazionale. La legge di conservazione afferma:

$$\sum_F w_F(S) \cdot I_F(S) = \Psi(S) = \text{costante rispetto alle trasformazioni che preservano la struttura auto-rappresentazionale}$$

dove  $w_F$  è il peso della famiglia  $F$  nel sistema  $S$ ,  $I_F$  è l'intensità dell'impossibilità di tipo  $F$ , e la somma è sulle cinque famiglie.

Questa legge non è ancora dimostrata in forma rigorosa — è una congettura strutturale che le analisi di questo ciclo rendono plausibile. Ma ha supporto empirico nei casi biologici: i sistemi biologici che riducono le impossibilità dimensionali (attraverso la sparsità e la struttura di varietà) aumentano necessariamente le impossibilità di irreversibilità (la sparsità richiede periodi di apprendimento intenso con successiva cristallizzazione, un processo irreversibile). I sistemi che riducono le impossibilità di complessità (attraverso le euristiche) aumentano le impossibilità di chiusura semantica (le euristiche sono accurate sintatticamente ma semanticamente distorte nei domini fuori dall'EEA).

La legge di conservazione dell'impossibilità, se dimostrata in forma rigorosa, costituirebbe il risultato fondamentale di questo ciclo: non soltanto le impossibilità sono necessarie, ma la loro quantità totale è conservata. I sistemi non possono sfuggire alle impossibilità — possono soltanto trasformarle, convertire un tipo in un altro, distribuire il peso dell'impossibilità tra le famiglie in modo diverso a seconda della propria architettura.

## **X. Perché le famiglie sono apparse storicamente come risultati separati**

La domanda inversa — perché strutture che sono corollari di un principio unico sono apparse come risultati separati in discipline separate per decenni — merita una risposta che non sia soltanto sociologica ma epistemologicamente illuminante.

La risposta è che le cinque famiglie si manifestano a scale e in contesti diversi che rendevano difficile riconoscerne la struttura comune. Il teorema di Gödel è un risultato della logica matematica, articolato nel linguaggio dei sistemi formali e delle funzioni ricorsive. Il teorema di Minsky-Papert è un risultato della teoria della computazione e dell'apprendimento automatico, articolato nel linguaggio della geometria euclidea e della separabilità lineare. Il teorema No Free Lunch è un risultato dell'ottimizzazione, articolato nel linguaggio degli algoritmi di ricerca e delle distribuzioni di probabilità. Il divario esplicativo è un argomento della filosofia della mente, articolato nel linguaggio dell'ontologia e dell'epistemologia. Il ratchet di Müller è un risultato della genetica delle popolazioni, articolato nel linguaggio della biologia evolutiva.

Ognuno di questi risultati è stato sviluppato da comunità di ricercatori che non leggevano le pubblicazioni delle altre comunità, che usavano notazioni radicalmente diverse, e che avevano obiettivi intellettuali distinti. Il riconoscimento della struttura comune richiede un livello di astrazione che trascende ciascuna disciplina — il livello della teoria delle categorie, della teoria dell'informazione, e della logica matematica astratta — e questo livello non era disponibile in nessuna delle discipline originarie al momento delle scoperte.

Ma la domanda più profonda è: perché la struttura comune non è stata riconosciuta nemmeno dopo la disponibilità degli strumenti astratti appropriati? Il teorema di Lawvere, che unifica le diagonalizzazioni in linguaggio categoriale, è del 1969 — ma la sua connessione con il binding problem neurobiologico o con il ratchet di Müller non è stata sviluppata nella letteratura. Il principio di Landauer, che connette la cancellazione informatica all'irreversibilità termodinamica, è del 1961 — ma la sua connessione con la chiusura dei periodi critici o con la perdita della semantica nell'apprendimento culturale non è stata sistematicamente esplorata.

La risposta è che la frammentazione disciplinare non è soltanto sociologica — è essa stessa una manifestazione della Famiglia III. Riconoscere la struttura comune tra risultati in domini diversi è un problema di complessità computazionale elevata: richiede la costruzione di mappe di riduzione tra strutture formali molto diverse, la verifica che queste mappe preservino le proprietà rilevanti, e la sintesi di queste mappe in un quadro coerente. Questo è un problema di tipo  $\Sigma_2^P$  o superiore nella gerarchia polinomiale — esattamente il tipo di problema che i sistemi cognitivi biologici gestiscono attraverso euristiche e approssimazioni invece dell'ottimizzazione esatta.

La frammentazione disciplinare della conoscenza scientifica è quindi in parte una risposta adattiva alle impossibilità di complessità del lavoro di sintesi: è più efficiente in termini computazionali lavorare all'interno di un dominio con le sue notazioni e metodi consolidati che attraversare continuamente i confini tra domini. Il costo è precisamente quello che questo ciclo si propone di recuperare: la perdita delle strutture unificanti che soltanto la visione trasversale rende visibili.

## **XI. Il PAI e la struttura della conoscenza scientifica**

Il principio di auto-rappresentazione insufficiente ha un'implicazione finale che porta questo ciclo nella filosofia della scienza: ogni sistema di conoscenza sufficientemente ricco — ogni disciplina scientifica, ogni sistema formale, ogni teoria del tutto — è soggetto al PAI. Nessuna teoria può essere simultaneamente consistente, completa, computazionalmente trattabile e informativamente conservativa rispetto al dominio che descrive.

Questo non è pessimismo epistemologico — è la struttura della conoscenza scientifica reale. La fisica quantistica è consistente e computazionalmente trattabile (per i sistemi a cui si applica) ma incompleta (il problema della misurazione) e informativamente non conservativa nella transizione classico-quantistico. La biologia evolutiva è consistente e semanticamente ricca ma computazionalmente intrattabile per la previsione di traiettorie evolutive specifiche. La neuroscienza è computazionalmente trattabile per certi domini ma incontra il divario esplicativo come impossibilità di chiusura semantica.

Non esistono teorie scientifiche che si sottraggano al PAI — esistono soltanto teorie che distribuiscono le impossibilità in modi diversi. La comprensione di

questa distribuzione — sapere quale famiglia di impossibilità è incontrata da quale teoria e perché — è la competenza epistemologica che una teoria unificata delle impossibilità strutturali può fornire. Non per sfuggire alle impossibilità, ma per abitarle consapevolmente: sapere esattamente dove si trova il confine è la condizione per muoversi il più vicino possibile ad esso senza la perdita di orientamento che la sua non-riconoscenza inevitabilmente produce.

---

## Capitolo VIII

### Il teorema del punto fisso come struttura trasversale

#### La grammatica degli invarianti, l'auto-riferimento produttivo e la topologia dell'impossibile

##### Mappa epistemica — Capitolo VIII

| Status          | Cod. | Claim                                                                                                                    |
|-----------------|------|--------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | Teoremi del punto fisso di Brouwer, Banach, Tarski (reticoli), Kleene, Schauder: tutti formalmente dimostrati.           |
| DIMOSTRATO      | D    | PPAD-completezza del Nash equilibrium (Chen-Deng-Teng 2009).                                                             |
| CONGETTURA      | C    | Il teorema di Banach come struttura formale della convergenza percettiva nel predictive coding: congettura ben fondata.  |
| INTERPRETAZIONE | I    | Il Cogito cartesiano come proposizione di punto fisso; i punti fissi dell'evoluzione culturale: proposte interpretative. |

#### I. Un filo che attraversa tutto

Il percorso di questo ciclo ha attraversato logica matematica, biologia evolutiva, neurobiologia cognitiva, geometria degli spazi metrici, teoria della complessità computazionale, termodinamica dell'informazione. In ciascuno di questi domini ha identificato strutture di impossibilità che condividono, secondo il capitolo VII, un principio generativo unico. Ma c'è un oggetto matematico — un tipo specifico di risultato — che compare in ciascuno di questi domini con tale regolarità da meritare un'analisi dedicata: il teorema del punto fisso.

Un punto fisso è un elemento  $x$  di uno spazio  $X$  che è invariante rispetto a una funzione  $f: X \rightarrow X$ , nel senso che  $f(x) = x$ . La semplicità dell'idea contrasta con la profondità delle sue implicazioni: i teoremi del punto fisso sono tra i risultati matematici più potenti e più universali, e la loro presenza nei contesti delle impossibilità strutturali non è accidentale. questo capitolo argomenta che i teoremi del punto fisso sono la grammatica delle impossibilità strutturali — la struttura formale che connette il PAI del capitolo VII alla fenomenologia concreta delle famiglie di impossibilità.

Il filo conduttore è la seguente osservazione: ogni impossibilità strutturale si manifesta come l'esistenza di un punto fisso indesiderato — un elemento invariante che il sistema non può né rimuovere né ignorare. Il teorema di Gödel usa la sentenza auto-referenziale  $G$  come punto fisso della funzione di negazione della dimostrabilità. Il teorema di Turing usa il programma diagonale  $D$  come punto fisso dell'operazione di commutazione  $HALT/LOOP$ . Il binding problem usa il pattern di interferenza come punto fisso dell'operazione di recovery nella sovrapposizione. Il divario esplicativo usa il quale come punto fisso che la fisica non può rimuovere senza contraddire la propria semantica fenomenica.

La prospettiva del punto fisso non è soltanto un modo alternativo di descrivere le impossibilità già note — è un modo che rivela connessioni nuove tra i teoremi del punto fisso della matematica pura (Brouwer, Banach, Knaster-Tarski, Kleene, Schauder) e le impossibilità strutturali della cognizione e della biologia. Questa rivelazione costituisce il contributo centrale di questo capitolo.

## II. La zoologia dei teoremi del punto fisso

Prima di sviluppare la tesi, è necessario costruire un catalogo abbastanza ricco dei teoremi del punto fisso per rendere visibile la loro ubiquità e per identificare le loro strutture formali distinte.

Il teorema di punto fisso di Brouwer (1910). Ogni funzione continua  $f: B^n \rightarrow B^n$  dalla palla unitaria chiusa in sé stessa ha almeno un punto fisso.

La dimostrazione classica usa argomenti topologici (la non-contrattibilità della sfera  $S^{n-1}$  come sottospazio di  $B^n$ ). Esiste anche una dimostrazione per via combinatoria (il lemma di Sperner). Il teorema di Brouwer è l'archetipo dei teoremi di esistenza: garantisce l'esistenza del punto fisso senza fornire un metodo per trovarlo. Ed è, in un senso preciso che verrà chiarito, computazionalmente irrisolvibile: trovare il punto fisso di una funzione continua generica su  $B^n$  è PPAD-completo (Papadimitriou 1994) — una classe di complessità che descrive problemi per cui l'esistenza di una soluzione è garantita da argomenti topologici o combinatoriali, ma trovare la soluzione richiede risorse superpolinomiali.

Il teorema di punto fisso di Banach (1922). Ogni funzione di contrazione  $f: (X, d) \rightarrow (X, d)$  su uno spazio metrico completo ha un unico punto fisso, raggiungibile iterando  $f$  a partire da qualsiasi punto iniziale.

Una funzione è una contrazione se esiste  $k \in [0,1)$  tale che  $d(f(x), f(y)) \leq k \cdot d(x, y)$  per tutti  $x, y \in X$ . Il punto fisso  $x^*$  è il limite della successione  $x_{n+1} = f(x_n)$ . La velocità di convergenza è geometrica:  $d(x_n, x^*) \leq k^n d(x_0, x^*)$ . Il teorema di Banach è costruttivo — fornisce sia l'esistenza sia un algoritmo efficiente per trovare il punto fisso.

Il teorema di Banach è il cuore della teoria delle iterazioni che descrive l'apprendimento per rinforzo (il TD learning del capitolo XIV del ciclo biologico è la realizzazione biologica dell'iterazione di Banach sull'operatore di Bellman), il predictive coding (la minimizzazione di  $F$  per iterazioni successive è una contrazione sotto certe condizioni), e la teoria delle frattali (gli iterated function systems di Barnsley).

Il teorema di punto fisso di Tarski (1955). Sia  $(L, \leq)$  un reticolo completo e sia  $f: L \rightarrow L$  una funzione monotona crescente ( $x \leq y \rightarrow f(x) \leq f(y)$ ). Allora  $f$  ha almeno un punto fisso, e l'insieme dei punti fissi di  $f$  forma a sua volta un reticolo completo.

Il teorema di Tarski è fondamentale nella teoria dei tipi, nella semantica dei linguaggi di programmazione (i punti fissi dei funtori di tipo sono i tipi ricorsivi) e nella teoria degli ordini. Il suo legame con le impossibilità strutturali è meno immediato ma profondo: il reticolo dei punti fissi di un operatore semantico è precisamente lo spazio delle "interpretazioni stabili" di un sistema formale — le strutture che il sistema non può modificare attraverso le proprie regole di inferenza.

Il teorema del punto fisso di Kleene (1938). Per qualsiasi funzione ricorsiva totale  $f: \mathbb{N} \rightarrow \mathbb{N}$ , esiste un numero  $e^*$  tale che  $\varphi_{f(e^*)} = \varphi_{e^*}$  — il programma con indice  $e^*$  ha lo stesso comportamento del programma con indice  $f(e^*)$ .

Il teorema di Kleene è il fondamento della semantica dei linguaggi di programmazione con ricorsione: il punto fisso dell'operatore funzionale  $F$  che definisce una funzione ricorsiva è la funzione stessa. Ogni definizione ricorsiva  $f(x)$

$= h(f, x)$  ha come punto fisso la funzione  $f$  — il punto fisso dell'operatore  $T_h: (\mathbb{N} \rightarrow \mathbb{N}) \rightarrow (\mathbb{N} \rightarrow \mathbb{N})$  definito da  $T_h(g)(x) = h(g, x)$ .

Il teorema di Kleene è strettamente connesso al lemma della diagonalizzazione di Gödel e ha la stessa struttura del teorema di Lawvere: è la versione computazionale del fatto che ogni operatore su funzioni ha un punto fisso.

Il teorema del punto fisso di Schauder (1930). Ogni funzione continua  $f: C \rightarrow C$  su un sottoinsieme convesso compatto  $C$  di uno spazio di Banach ha almeno un punto fisso.

Il teorema di Schauder è una generalizzazione del teorema di Brouwer agli spazi di dimensione infinita — cruciale per la teoria delle equazioni differenziali funzionali, per la teoria dei giochi con spazi di strategie infinito-dimensionali, e per la fisica matematica. La sua connessione con le impossibilità strutturali biologiche passa attraverso i sistemi dinamici in spazi funzionali: l'evoluzione di un sistema cognitivo può essere modellata come un flusso su uno spazio di Banach di distribuzioni di probabilità, e il teorema di Schauder garantisce l'esistenza di stati stazionari (attractor states) che il sistema può raggiungere attraverso l'apprendimento.

### **III. Il teorema di Brouwer e la PPAD-completezza: l'impossibilità costruttiva**

La connessione tra il teorema di Brouwer e le impossibilità di complessità è particolarmente istruttiva perché rivela un aspetto delle impossibilità strutturali che le analisi precedenti non avevano tematizzato: la differenza tra l'esistenza di una soluzione e la sua trovabilità efficiente.

Papadimitriou (1994) ha identificato la classe di complessità PPAD (Polynomial Parity Argument on Directed graphs) come la classe dei problemi per cui l'esistenza di una soluzione è garantita da argomenti del tipo "ogni grafo orientato ha un numero pari di nodi di grado dispari" — argomenti che non costruiscono la soluzione ma ne garantiscono l'esistenza per ragioni combinatoriali o topologiche. Il problema di trovare il punto fisso di una funzione continua in  $B^n$  (nella versione discreta appropriata) è PPAD-completo.

Teorema (Chen, Deng, Teng 2009). Trovare un equilibrio di Nash  $\epsilon$ -approssimato in un gioco bimatriciale è PPAD-completo.

Poiché ogni equilibrio di Nash è un punto fisso della corrispondenza di best-response (ciascun giocatore risponde ottimalmente alla strategia degli avversari, e all'equilibrio nessuno vuole deviare — ciascuna strategia è il proprio "punto fisso" rispetto alle best-responses), la PPAD-completezza del Nash equilibrium è la PPAD-completezza del problema del punto fisso di Brouwer per la corrispondenza di best-response.

La connessione con le impossibilità biologiche è diretta: il capitolo XIV del ciclo biologico ha analizzato l'evoluzione della cooperazione come teoria dei giochi evolutivi, e gli equilibri evolutivi stabili (ESS) sono i punti fissi della dinamica replicatrice. La PPAD-completezza del problema di trovare questi equilibri implica che i sistemi biologici non possono trovare gli equilibri evolutivi ottimali in tempo polinomiale — devono convergere a equilibri sub-ottimali attraverso processi heuristici (la selezione naturale come algoritmo di discesa sul gradiente dell'idoneità) che non garantiscono l'ottimalità globale.

Questo è un esempio preciso della struttura del PAI applicato ai sistemi evolutivi: la selezione naturale produce sistemi che si avvicinano agli equilibri evolutivi ma non li raggiungono esattamente, perché il raggiungimento esatto richiederebbe risorse computazionali PPAD-complete — risorse che eccedono qualsiasi sistema biologico fisicamente realizzabile.

#### **IV. Il teorema di Banach e la contrazione cognitiva**

Il teorema di Banach — il teorema del punto fisso per le contrazioni sugli spazi metrici completi — ha una presenza pervasiva nelle strutture cognitive biologiche che la letteratura neuroscientifica ha riconosciuto soltanto parzialmente.

Il principio di energia libera di Friston (Capitolo VIII del ciclo biologico) è formalmente la minimizzazione di  $F$  attraverso una sequenza di aggiornamenti. Per il framework del predictive coding, l'aggiornamento delle credenze è:

$$\mu_{+1} = \mu - \epsilon \cdot \partial F / \partial \mu$$

dove  $\mu$  è il vettore dei parametri del modello generativo e  $\epsilon$  è il passo di aggiornamento. Questo è un passo del metodo del gradiente, che converge al

minimo di  $F$  se  $F$  è convessa e lo step-size  $\epsilon$  è sufficientemente piccolo. Ma la convergenza del gradient descent è garantita (con velocità geometrica) esattamente quando l'operatore di aggiornamento  $T(\mu) = \mu - \epsilon \cdot \partial F / \partial \mu$  è una contrazione di Banach:

$$\|T(\mu_1) - T(\mu_2)\| \leq k \cdot \|\mu_1 - \mu_2\| \quad \text{con } k < 1$$

La condizione di contrazione richiede che la Hessiana di  $F$  abbia autovalori limitati superiormente: se  $\lambda_{\max}$  è il massimo autovalore di  $\partial^2 F / \partial \mu^2$ , la contrazione vale per  $\epsilon \leq 2/\lambda_{\max}$ . In questo regime, il sistema cognitivo converge geometricamente al minimo di  $F$  — il punto fisso del processo di apprendimento.

Il teorema di Banach ha quindi una lettura cognitiva precisa: il punto fisso del processo di minimizzazione di  $F$  è lo "stato percettivo stabile" — la credenza stabile che il sistema cognitivo mantiene quando i segnali sensoriali sono coerenti con le previsioni del modello. La convergenza a questo punto fisso è il processo della percezione: vedere stabilmente un oggetto, udire stabilmente una parola, riconoscere stabilmente un pattern.

Ma la condizione di contrazione introduce un nuovo tipo di impossibilità: quando  $F$  non è convessa — quando il paecapitolo dell'energia libera ha più minimi locali — l'operatore  $T$  non è una contrazione globale, e il sistema converge a un minimo locale dipendente dalle condizioni iniziali (path dependence della Famiglia V). La ricchezza della vita cognitiva — la molteplicità delle percezioni possibili, le illusioni, i cambiamenti di prospettiva — è precisamente la manifestazione del fatto che  $F$  ha tipicamente molti minimi locali, e la convergenza a ciascuno di essi è una realizzazione diversa del teorema di Banach in regioni diverse dello spazio dei parametri.

Struttura formale dei minimi locali come punti fissi multipli. L'insieme dei punti fissi di  $T$  in presenza di minimi locali multipli ha la struttura di una varietà discreta — un insieme di punti isolati nel parametri-spazio. Ciascun punto fisso è un "attrattore" percettivo stabile. Le illusioni percettive (il cubo di Necker, la maschera concava, la figura Rubin del vaso/volti) sono sistemi in cui il paecapitolo  $F$  ha due o più minimi locali di energia comparabile, e il sistema oscilla tra di essi o si stabilizza in uno a seconda della condizione iniziale (l'attenzione, il contesto, la storia recente dell'osservazione). Queste illusioni non sono difetti della percezione

— sono manifestazioni dirette della struttura dei punti fissi multipli del processo di minimizzazione di  $F$ .

## V. Il teorema di Tarski-Kleene e la semantica dei linguaggi ricorsivi

Il teorema di punto fisso di Tarski (nella versione per reticoli) e il teorema di Kleene (nella versione per le funzioni ricorsive) sono entrambi fondativi per la semantica dei linguaggi di programmazione con strutture ricorsive — e questa semantica ha connessioni profonde con la struttura dell'impossibilità di chiusura semantica (Famiglia IV).

Un linguaggio di programmazione con tipi ricorsivi — tipi che possono contenere istanze di se stessi, come le liste (una lista è vuota o è un elemento seguito da una lista) o gli alberi — richiede la soluzione di equazioni di tipo della forma:

$$T = F(T)$$

dove  $F$  è un funtore che costruisce nuovi tipi da tipi esistenti. La soluzione di questa equazione è il punto fisso del funtore  $F$  nell'appropriato reticolo dei tipi. Per i tipi ricorsivi degli ordini di Haskell o dei tipi dipendenti di Coq, il punto fisso del funtore è costruito attraverso il teorema di Tarski applicato al reticolo dei tipi ordinati per inclusione.

La connessione con la Famiglia IV è la seguente: la definizione di tipi ricorsivi attraverso equazioni di punto fisso è un esempio controllato di auto-riferimento — un caso in cui il sistema può fare riferimento a se stesso senza generare contraddizione, perché la ricorsione è "guardata" (well-founded) nel senso che ogni chiamata ricorsiva riduce la dimensione dell'input. I linguaggi con tipi ricorsivi guardate sono consistenti e completi per i tipi che esprimono.

Ma esistono tipi ricorsivi non guardate — tipi definiti da equazioni di punto fisso in cui la ricorsione non è ben fondata. Il tipo  $T = T \rightarrow \text{Bool}$  (un tipo uguale a se stesso applicato come funzione a Bool) non è ben fondato — porta a inconsistenze nel sistema di tipi. Questo è esattamente il caso in cui il teorema di Tarski non garantisce l'esistenza del punto fisso nel reticolo dei tipi ordinati, e la soluzione formale richiederebbe estendere il reticolo a un dominio di Scott — una

struttura topologica che permette di dare significato a calcoli "parzialmente definiti" a costo di abbandonare la completezza.

Il calcolo- $\lambda$  non tipato (in cui qualsiasi termine può essere applicato a qualsiasi altro) è il sistema che rende tutto ciò massimamente esplicito: il termine  $Y = \lambda f.(\lambda x.f(xx))(\lambda x.f(xx))$  è un combinatore del punto fisso che soddisfa  $Yf = f(Yf)$  per qualsiasi  $f$  — esiste un punto fisso per ogni funzione  $f$  nel calcolo  $\lambda$ . Ma questa onnipresenza dei punti fissi si ottiene abbandonando la consistenza tipistica: il calcolo- $\lambda$  non tipato è inconsistente come teoria logica (può dimostrare qualsiasi proposizione). Il punto fisso universale  $Y$  è la manifestazione del PAI nella sua forma più pura: l'universalità dei punti fissi implica l'inconsistenza del sistema.

## **VI. I punti fissi nell'evoluzione: attrattori evolutivi e strutture stabili**

I punti fissi nei sistemi evolutivi hanno una struttura che la biologia teorica ha studiato sotto il nome di "equilibri evolutivamente stabili" (ESS) — un concetto introdotto da Maynard Smith e Price (1973) come generalizzazione del concetto di strategia ottimale per sistemi in cui l'idoneità di una strategia dipende dalla frequenza degli avversari.

Definizione (equilibrio evolutivamente stabile, ESS). Una strategia  $s^*$  è un ESS se per ogni strategia mutante  $s \neq s^*$ :

$$W(s^*, s^*) > W(s, s^*), \text{ oppure } W(s^*, s^*) = W(s, s^*) \text{ e } W(s^*, s) > W(s, s)$$

dove  $W(a, b)$  è l'idoneità della strategia  $a$  quando giocata contro la strategia  $b$ .

Un ESS è un punto fisso della dinamica replicatrice — lo stato della popolazione che rimane invariato sotto la selezione naturale in assenza di mutazioni. La condizione di ESS garantisce la stabilità del punto fisso contro le invasioni delle strategie mutanti.

La struttura dei punti fissi ESS nell'evoluzione ha proprietà notevoli:

Non tutti i giochi hanno ESS: il gioco Pietra-Carta-Forbice (RPS) non ha ESS in strategie pure — la dinamica replicatrice orbita attorno all'equilibrio misto senza convergere. L'equilibrio misto  $(1/3, 1/3, 1/3)$  è un punto fisso nel senso

matematico, ma è neutralmente stabile invece che evolutivamente stabile — una perturbazione non lo destabilizza, ma il sistema non ritorna attivamente ad esso.

I giochi coevolutivi (in cui l'idoneità di ciascuna specie dipende dalla composizione dell'altra) producono dinamiche di punti fissi multipli con struttura di biforcazione: al variare dei parametri del gioco, il numero e la stabilità degli ESS cambia qualitativamente. Questi cambiamenti qualitativo — le biforcazioni — corrispondono alle transizioni evolutive discusse nel capitolo VII del ciclo biologico.

La connessione con le impossibilità strutturali è la PPAD-completezza già discussa: trovare un ESS o un equilibrio di Nash  $\epsilon$ -approssimato è PPAD-completo. Questo significa che l'evoluzione non "trova" l'ESS attraverso un algoritmo efficiente — converge a soluzioni approssimate che dipendono dalla storia evolutiva (path dependence della Famiglia V) e che possono essere distanti dall'ESS esatto.

## **VII. Il punto fisso dell'auto-conoscenza: il Cogito come proposizione di punto fisso**

Un'applicazione dei teoremi del punto fisso alla filosofia della mente che non è stata sviluppata nella letteratura è la riformulazione del Cogito cartesiano — "cogito ergo sum" — come proposizione di punto fisso.

Il Cogito afferma che il dubbio su ogni proposizione è impossibile per la proposizione "io esisto come pensante": l'atto di dubitare del proprio pensiero è esso stesso un pensiero, il che dimostra l'esistenza del pensiero. In termini di punto fisso: sia  $f$  l'operazione di dubbio su una proposizione. La proposizione "io penso" è un punto fisso di  $f$  nel senso che  $f(\text{"io penso"}) = \text{"io penso"}$  — dubitare del proprio pensiero è ancora pensare, e quindi la proposizione è invariante rispetto all'operazione di dubbio.

Questo non è soltanto una riformulazione elegante del Cogito — è una riformulazione che chiarisce la struttura dell'argomento e ne mostra sia la forza sia i limiti. La forza: il Cogito identifica una proposizione genuinamente auto-referenzialmente stabile — un punto fisso della funzione di dubbio che non è generato artificialmente ma emerge dalla struttura stessa del pensiero. Il limite: la

stabilità rispetto al dubbio non è la stessa cosa della verità — un punto fisso non è necessariamente un punto di verità. Il Cogito dimostra che "io penso" è immune al dubbio (è un punto fisso della funzione di dubbio), non che "io penso" sia una proposizione sulla natura della sostanza pensante.

La riformulazione in termini di punto fisso connette il Cogito alla struttura della Famiglia IV: la proposizione "io penso" è l'unica proposizione del sistema cognitivo che è un punto fisso della propria funzione semantica — l'unica proposizione che non può essere negata senza auto-contraddizione. È il fondo roccioso dell'auto-riferimento semantico — il punto al di sotto del quale nessun sistema cognitivo può scendere senza dissolversi.

La connessione con il teorema di Kleene è la seguente: nella terminologia computazionale, il "programma" che implementa il pensiero è il proprio punto fisso nella numerazione di Kleene — il programma  $e^*$  tale che  $\varphi_{-}\{e^*\} = \varphi_{-}\{e^*\}$ . Il Cogito è la versione filosofica del teorema di ricorsione di Kleene: esiste sempre un punto dell'auto-riferimento computazionale in cui il sistema rimane invariato rispetto alla propria auto-applicazione.

## **VIII. Il punto fisso dell'evoluzione culturale: la lingua come attrattore**

il capitolo VII del ciclo biologico ha analizzato la trasmissione culturale come sistema soggetto a derive e perdite. Ma la prospettiva del punto fisso rivela un aspetto complementare: le strutture culturali che sopravvivono nel lungo periodo non sono strutture qualsiasi — sono strutture che sono punti fissi (o attrattori) del processo di trasmissione culturale.

Il modello formale di Bartlett — la trasmissione di storie come processo iterativo  $\tilde{s}_i = A \cdot \tilde{s}_{i-1} + b$  — ha il punto fisso  $s^* = (I - A)^{-1}b$ . Le strutture che convergono a questo punto fisso sono precisamente i "prototipi culturali" — le storie, i miti, i rituali che resistono alla deriva perché corrispondono al punto fisso dell'operatore di trasmissione. Non è che le storie vengano conservate perché sono vere o belle o utili — vengono conservate perché sono punti fissi del processo cognitivo di trasmissione. La verità, la bellezza e l'utilità possono coincidere con la proprietà del punto fisso, ma non è necessario.

Questa prospettiva ha conseguenze per la comprensione dell'evoluzione delle lingue naturali. Kirby, Cornish e Smith (2008) hanno mostrato sperimentalmente che le lingue artificiali trasmesse per iterazioni convergono a strutture composizionali e regolari. Nella prospettiva del punto fisso: le lingue naturali sono i punti fissi dell'operatore di trasmissione linguistica — le strutture che la mente umana può apprendere, usare e trasmettere senza degradazione progressiva.

La struttura di una lingua naturale è quindi il punto fisso di un sistema di equazioni funzionali che vincolano simultaneamente:

- (a) L'apprendibilità: la struttura deve essere acquisibile da bambini in pochi anni attraverso esposizione implicita (il periodo critico linguistico).
- (b) La trasmissibilità: la struttura deve essere trasmissibile da adulto a adulto con bassa degenerazione.
- (c) L'usabilità: la struttura deve permettere la generazione e la comprensione di enunciati nuovi in tempo reale.
- (d) L'espressività: la struttura deve essere sufficientemente ricca da esprimere le proposizioni cognitivamente rilevanti per l'ambiente dell'utente.

Queste quattro condizioni sono vincoli funzionali sull'operatore di trasmissione linguistica, e il punto fisso di questo operatore — la struttura che soddisfa simultaneamente tutte e quattro le condizioni — è la grammatica universale nel senso di Chomsky: non un insieme di regole innate, ma una struttura di attrazione nello spazio delle lingue possibili verso cui qualsiasi processo di evoluzione linguistica converge.

## **IX. La topologia dei punti fissi: varietà stabili e instabili**

I punti fissi di un sistema dinamico non sono tutti uguali: alcuni sono stabili (i punti fissi attrattori, verso cui il sistema converge quando perturbato), altri sono instabili (i punti fissi repulsori, da cui il sistema si allontana quando perturbato), altri ancora sono punti di sella (attrattori lungo certe direzioni, repulsori lungo altre).

La struttura locale di un punto fisso  $x^*$  di un sistema dinamico  $\dot{x} = f(x)$  è determinata dagli autovalori della matrice jacobiana  $J = \partial f / \partial x |_{x^*}$ :

- Tutti gli autovalori con parte reale negativa:  $x^*$  è un attrattore stabile — le perturbazioni si estinguono esponenzialmente.
- Tutti gli autovalori con parte reale positiva:  $x^*$  è un repulsore instabile — le perturbazioni crescono esponenzialmente.
- Autovalori con parte reale di segno misto:  $x^*$  è un punto di sella — il sistema converge lungo le varietà stabili (corrispondenti agli autovettori con autovalori negativi) e diverge lungo le varietà instabili (corrispondenti agli autovettori con autovalori positivi).

Questa struttura topologica dei punti fissi è direttamente rilevante per le impossibilità strutturali.

Le impossibilità diagonali corrispondono a punti fissi repulsori: la sentenza gödeliana  $G$  è il punto fisso della funzione di negazione della dimostrabilità, e  $G$  "respinge" qualsiasi tentativo di decidere il suo valore di verità — il sistema si allontana da  $G$  invece di convergere ad essa.

Le impossibilità dimensionali corrispondono a punti di sella: la proiezione ottimale (il punto fisso dell'algoritmo di Johnson-Lindenstrauss) è stabile nella direzione delle dimensioni preservate (il sistema converge rapidamente alla proiezione ottimale lungo quelle dimensioni) e instabile nella direzione delle dimensioni perse (qualsiasi perturbazione lungo le dimensioni compresse produce amplificazione della distorsione).

Le impossibilità di irreversibilità corrispondono a punti fissi attrattori con bacini di attrazione separati da barriere: gli stati terminali dei processi irreversibili (i tipi cellulari dopo la differenziazione, la struttura neurale dopo la chiusura del periodo critico) sono attrattori stabili con barriere alte che impediscono le transizioni verso altri attrattori.

Questa topologia dei punti fissi fornisce una mappa geometrica dello spazio delle impossibilità: i punti fissi repulsori sono le impossibilità che il sistema cerca attivamente di evitare; i punti di sella sono le impossibilità che il sistema può aggirarsi con cura ma non eliminare; gli attrattori con alte barriere sono le impossibilità che il sistema incontra definitivamente una volta entrato nel loro bacino di attrazione.

## X. Il punto fisso come struttura della creatività

Questo ciclo si è costruito attorno alla tesi che le impossibilità strutturali sono la condizione della possibilità del pensiero — che i confini siano mappe, non muri. Il punto fisso è il luogo in cui questa tesi si fa più concreta.

I punti fissi stabili — gli attrattori — sono le strutture invarianti che permettono l'identità nel tempo: la percezione stabile di un oggetto, l'identità di una lingua, l'equilibrio di un ecosistema, la struttura di un tipo cellulare. Senza punti fissi stabili non vi è identità, non vi è continuità, non vi è possibilità di costruire su ciò che si è già costruito. I punti fissi stabili sono la base della conoscenza accumulata.

I punti fissi instabili — i repulsori — sono le strutture che il sistema non può mantenere: le credenze contraddittorie, le rappresentazioni sovra sature, le strategie evolutivamente instabili. I repulsori non sono ostacoli — sono frecce direzionali che indicano dove il sistema non deve rimanere, e quindi indicano implicitamente dove deve andare. L'impossibilità diagonale — la sentenza gödeliana G che è un punto fisso repulsore — non blocca il pensiero: lo respinge oltre il sistema corrente, verso il livello metalinguistico superiore in cui G diventa decidibile.

I punti di sella sono i più interessanti dal punto di vista della creatività: sono strutture che sono stabili in alcune direzioni e instabili in altre. Lungo le direzioni instabili, il sistema diverge rapidamente — piccole perturbazioni producono grandi cambiamenti. Questa sensibilità alle perturbazioni lungo le direzioni instabili è il meccanismo della creatività: una piccola variazione nella configurazione iniziale produce un percorso evolutivo radicalmente diverso. L'insight (la biforcazione termodinamica del capitolo XII del ciclo biologico) è il pascapitolo repentino da un bacino di attrazione a un altro — una traiettoria che passa vicino a un punto di sella e devia lungo la varietà instabile verso un nuovo attrattore.

Il teorema del punto fisso come struttura trasversale di tutte le impossibilità strutturali è quindi anche la struttura trasversale di tutta la creatività: creare è trovare nuovi punti fissi — nuove strutture invarianti che il sistema non aveva ancora raggiunto — attraverso la navigazione del paecapitolo dei punti fissi stabili, instabili e di sella che le impossibilità strutturali delineano. Ogni impossibilità strutturale è anche la struttura di una possibilità non ancora realizzata: il punto

fisso repulsore indica la direzione verso il livello superiore; il punto di sella indica il percorso attraverso la transizione creativa; l'attrattore indica la struttura stabile che attende alla fine della transizione.

Il confine dell'impossibile è la mappa della creatività.

## **XI. La grammatica trasversale: sintesi**

Le otto sezioni precedenti hanno sviluppato una grammatica dei punti fissi che attraversa le cinque famiglie di impossibilità strutturali:

Famiglia I (diagonale): il punto fisso repulsore della funzione di auto-riferimento; la sentenza di Gödel, il programma di Turing, il paradosso di Russell sono punti fissi che il sistema non può né raggiungere né evitare — il sistema orbita attorno a essi senza stabilizzarsi, ed è questa orbita instabile che produce l'incompletezza.

Famiglia II (dimensionale): il punto di sella della proiezione ottimale; la proiezione di Johnson-Lindenstrauss è stabile nelle dimensioni preservate e instabile nelle dimensioni perse — la distorsione è la manifestazione della varietà instabile del punto di sella.

Famiglia III (complessità): il punto fisso PPAD-completo della corrispondenza di best-response; trovare l'equilibrio di Nash è trovare il punto fisso della best-response, e la PPAD-completezza misura la difficoltà di questo compito.

Famiglia IV (chiusura semantica): l'impossibilità del punto fisso della funzione semantica; il teorema di Tarski mostra che la funzione di verità non ha punto fisso nel linguaggio oggetto — il tentativo di costruire il punto fisso genera contraddizione.

Famiglia V (irreversibilità): l'attrattore irraggiungibile dopo la biforcazione; i processi irreversibili convergono ad attrattori stabili con alte barriere, e l'irreversibilità è l'impossibilità di uscire dal bacino di attrazione raggiunto.

Questa grammatica è la sintesi formale dell'intero ciclo: le impossibilità strutturali non sono semplicemente "cose che non si possono fare" — sono strutture topologiche precise nello spazio dei sistemi auto-rappresentazionali. Comprendere quella topologia — i punti fissi, i loro tipi di stabilità, le loro connessioni, la

conservazione della loro carica impossibilitante — è comprendere la geometria del confine tra ciò che è pensabile e ciò che non lo è.

E come ogni geometria, questa topologia non è soltanto descrittiva — è normativa: dice dove si trovano i percorsi, dove le barriere, dove le transizioni possibili tra un regime e l'altro. Il confine dell'impossibile, mappato con sufficiente precisione, diventa la guida più affidabile che il pensiero possa avere per navigare il territorio in cui si muove.

---

## Capitolo IX

### Le impossibilità della scienza come sistema cognitivo

**Il problema della demarcazione, la struttura delle rivoluzioni,  
l'indeterminazione della traduzione e i limiti dell'inferenza  
induttiva**

#### Mappa epistemica — Capitolo IX

| Status          | Cod. | Claim                                                                                                                                                                                                                                |
|-----------------|------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | Struttura formale del teorema di Duhem-Quine: non esiste esperimento che falsifichi una singola ipotesi in isolamento.                                                                                                               |
| DIMOSTRATO      | D    | Complessità di campionamento PAC (Valiant 1984): $m(\epsilon, \delta) = \Theta(d_{VC}/\epsilon + \log(1/\delta)/\epsilon)$ .                                                                                                         |
| INTERPRETAZIONE | I    | Rivoluzioni di Kuhn come transizioni di fase; criterio di Popper come risposta all'impossibilità di chiusura semantica; indeterminazione di Quine come diagonalizzazione semantica: proposte interpretative strutturalmente fondate. |

#### I. La scienza come sistema auto-rappresentazionale

Gli otto capitoli che precedono hanno sviluppato una teoria delle impossibilità strutturali applicata a sistemi formali, sistemi biologici e sistemi cognitivi. Ma vi è un sistema cognitivo di importanza eccezionale che non è stato ancora analizzato in questa prospettiva: la scienza stessa — il sistema istituzionale e metodologico attraverso cui le comunità umane producono, validano e trasmettono conoscenza sul mondo.

La scienza non è semplicemente un insieme di teorie vere sul mondo: è un sistema auto-rappresentazionale nel senso preciso del PAI. È un sistema che possiede una struttura interna di rappresentazione (le teorie, i modelli, le ipotesi) che si propone di catturare completamente la struttura del dominio che descrive (la natura, la realtà). Come tale, il PAI si applica: la scienza incontra necessariamente almeno una delle cinque famiglie di impossibilità strutturali.

Non soltanto le singole teorie scientifiche sono soggette alle impossibilità strutturali — tutta la letteratura della filosofia della scienza del Novecento è sostanzialmente la documentazione sistematica di queste impossibilità. Popper ha documentato l'impossibilità di chiusura semantica (il problema della demarcazione come impossibilità di verificare definitivamente le teorie). Kuhn ha documentato l'irreversibilità (le rivoluzioni scientifiche come transizioni di fase irreversibili tra paradigmi). Quine ha documentato l'impossibilità diagonale nella linguistica (l'indeterminazione della traduzione come forma di diagonalizzazione semantica). Hume ha documentato l'impossibilità di complessità (il problema dell'induzione come intrattabilità dell'inferenza dal particolare all'universale).

questo capitolo rilegge questi risultati classici alla luce della teoria delle impossibilità strutturali sviluppata nei cicli precedenti, mostrando che non sono osservazioni indipendenti sulla fragilità della conoscenza scientifica ma manifestazioni coerenti di un unico principio — il PAI — applicato al sistema cognitivo collettivo della scienza.

## **II. Il problema della demarcazione di Popper come impossibilità di chiusura semantica**

Karl Popper propose il criterio di falsificabilità come soluzione al problema della demarcazione — la distinzione tra affermazioni scientifiche genuine (quelle che possono essere confutate dall'esperienza) e pseudoscienza o metafisica (quelle che non possono esserlo). Una teoria è scientifica se e soltanto se è falsificabile: se esiste almeno un possibile risultato sperimentale che, se osservato, la renderebbe falsa.

Il criterio di falsificabilità è, nella terminologia di questo ciclo, un tentativo di trovare il confine semantico della scienza — il confine tra le affermazioni che hanno contenuto empirico (che possono essere confrontate con la realtà) e quelle che non lo hanno. Ma questo confine è soggetto al PAI nella sua forma più diretta: la Famiglia IV.

Il problema della falsificabilità come impossibilità di chiusura semantica. Ogni teoria scientifica  $T$  è connessa all'evidenza empirica  $E$  attraverso una catena di connessioni che include ipotesi ausiliarie  $H_1, \dots, H_n$  — ipotesi sui sistemi di

misurazione, sulle condizioni sperimentali, sulle condizioni al contorno. La struttura logica del test empirico non è:

$T \rightarrow e$  (la teoria implica direttamente l'osservazione  $e$ )

ma:

$$T \wedge H_1 \wedge \dots \wedge H_n \rightarrow e$$

dove  $H_i$  sono le ipotesi ausiliarie. Se l'osservazione  $e$  non si verifica ( $\neg e$ ), il modus tollens dà:

$$\neg(T \wedge H_1 \wedge \dots \wedge H_n) = \neg T \vee \neg H_1 \vee \dots \vee \neg H_n$$

La falsificazione dell'osservazione  $e$  non determina quale componente del congiunzione  $T \wedge H_1 \wedge \dots \wedge H_n$  è falsa: potrebbe essere  $T$ , oppure qualsiasi  $H_i$ . Il sistema scientifico può sempre salvare  $T$  modificando le ipotesi ausiliarie — il fenomeno che Lakatos (1970) ha chiamato il "cintura protettiva" del programma di ricerca.

Questo è un caso preciso di impossibilità di chiusura semantica: la teoria  $T$  non può determinare, dall'interno del sistema scientifico (che include  $T$  più tutte le ipotesi ausiliarie), se è  $T$  ad essere falsa o se è qualche  $H_i$ . La determinazione di questo richiederebbe un meta-sistema che osservi il sistema  $T \wedge H$  dall'esterno — ma qualsiasi meta-sistema è esso stesso soggetto alla stessa struttura.

La formalizzazione usa il teorema di Duhem-Quine: non esiste alcun esperimento che possa falsificare una singola ipotesi in isolamento; ogni test empirico è un test di sistemi di ipotesi. In termini del gap informativo del capitolo III:

$$\text{Gap}(T, \text{Evidenza}) = H(T_{\text{vera}}) - I(T_{\text{vera}}; \text{Evidenza})$$

è sempre positivo: la verità di  $T$  non è completamente determinata dall'evidenza empirica disponibile, perché la relazione tra  $T$  e l'evidenza è mediata dalle ipotesi ausiliarie che non sono specificate dall'evidenza stessa.

Il criterio di Popper come regola procedurale, non semantica. La risoluzione pratica che Popper adotta — trattare la falsificabilità come un criterio metodologico (una regola su come la scienza deve procedere) invece che come un criterio semantico (una proprietà delle teorie stesse) — è precisamente il tipo di spostamento di livello che il PAI prevede. Quando la chiusura semantica al livello

della teoria è impossibile, si sale al livello del metodo — si costruisce una norma procedurale (le teorie falsificabili devono essere abbandonate quando falsificate) che compensa la mancanza di chiusura semantica al livello della teoria.

Ma questo spostamento di livello introduce le impossibilità del livello superiore: la norma procedurale è essa stessa soggetta al PAI. Il criterio "abbandonare le teorie falsificate" è praticabile soltanto se si può determinare quando una teoria è stata genuinamente falsificata invece di essere "protetta" dalla modifica delle ipotesi ausiliarie. E questa determinazione richiede un giudizio qualitativo — il giudizio di "serendipità" di Lakatos, il "normale scienza" di Kuhn — che non è riducibile a un algoritmo.

### **III. Le rivoluzioni scientifiche di Kuhn come transizioni di fase irreversibili**

Thomas Kuhn (1962) propose un modello dello sviluppo scientifico che differisce radicalmente dal modello cumulativo di Popper: la scienza non avanza attraverso la falsificazione e la sostituzione graduale di teorie false con teorie migliori, ma attraverso rivoluzioni periodiche — transizioni discontinue tra "paradigmi" incompatibili.

Nella terminologia di questo ciclo, le rivoluzioni scientifiche di Kuhn sono transizioni di fase irreversibili nel senso della Famiglia V. Il paradigma scientifico dominante è un attrattore stabile del sistema cognitivo collettivo della comunità scientifica — uno stato verso cui il sistema converge durante il periodo di "scienza normale" e che resiste alle perturbazioni (le "anomalie") attraverso la modificazione delle ipotesi ausiliarie. La transizione tra paradigmi è una biforcazione irreversibile: il sistema passa da un bacino di attrazione (il vecchio paradigma) a un altro (il nuovo paradigma), e questa transizione non è invertibile attraverso un semplice recupero del vecchio paradigma.

La formalizzazione usa il modello del paesaggio energetico del capitolo VI. Sia  $U(p)$  il "potenziale epistemico" del paradigma  $p$  — una misura della coerenza interna del paradigma con l'evidenza accumulata. Durante la scienza normale,  $U(p_{\text{corrente}})$  è basso (il paradigma ha alta coerenza con l'evidenza) e il sistema cognitivo collettivo è in un minimo del paesaggio. Le anomalie — osservazioni

che il paradigma corrente non riesce a spiegare — aumentano localmente  $U(p_{\text{corrente}})$ : l'energia del minimo cresce, ma il sistema rimane nel bacino di attrazione del vecchio paradigma finché le barriere verso i minimi alternativi rimangono più alte dell'energia del sistema.

La crisi scientifica — il periodo in cui le anomalie si accumulano e il vecchio paradigma è messo in questione senza che un nuovo paradigma sia ancora disponibile — è il sistema che si avvicina alla barriera energetica tra i due bacini. La rivoluzione scientifica è la transizione: il sistema attraversa la barriera e crolla nel bacino del nuovo paradigma. L'irreversibilità è strutturale: una volta che la comunità scientifica ha adottato il nuovo paradigma, tornare al vecchio richiederebbe "disimparare" un insieme di strutture cognitive consolidate — il tipo di processo che le impossibilità della Famiglia V rendono fondamentalmente difficile.

L'incommensurabilità come impossibilità di chiusura semantica inter-paradigmatica. Kuhn introduce il concetto di "incommensurabilità" tra paradigmi: i termini teorici chiave del vecchio paradigma (massa, forza, elemento chimico, spazio, tempo) cambiano significato nel passaggio al nuovo paradigma, in modo tale che non è possibile tradurre completamente il vecchio paradigma nel linguaggio del nuovo. La "massa" nella meccanica classica e la "massa" nella relatività speciale non sono lo stesso concetto, anche se condividono molte proprietà formali.

L'incommensurabilità è una forma di impossibilità di chiusura semantica inter-paradigmatica: il linguaggio del vecchio paradigma non può catturare completamente la semantica del nuovo, perché le relazioni di soddisfazione tra i termini teorici e il dominio empirico cambiano qualitativamente nel passaggio da un paradigma all'altro. Il gap semantico tra i paradigmi è positivo nel senso del capitolo III:  $\text{Gap}(L_{\text{vecchio}}, \text{Semantica}_{\text{nuovo}}) > 0$ .

Questo produce il famoso problema della scelta razionale tra paradigmi: se i paradigmi sono incommensurabili, come può la comunità scientifica scegliere razionalmente tra di essi? La risposta di Kuhn — che la scelta è parzialmente extra-razionale, mediata da valori estetici, sociologici e generazionali — è la risposta che il PAI prevede: quando la chiusura semantica è impossibile al livello del confronto diretto tra paradigmi, il sistema cognitivo collettivo compensa con

meccanismi extra-semantici (la persuasione, la sociologia della scienza, il ricambio generazionale — "la scienza avanza un funerale alla volta", nella sarcastica osservazione di Max Planck).

#### **IV. L'indeterminazione della traduzione di Quine come diagonalizzazione semantica**

Willard Van Orman Quine (1960) propose l'argomento dell'indeterminazione della traduzione: per qualsiasi lingua sconosciuta L e qualsiasi corpus di comportamento linguistico dei parlanti di L, esistono più manuali di traduzione da L a una lingua di riferimento R che sono tutti ugualmente compatibili con il comportamento osservato ma che si contraddicono nell'assegnare significati a certi enunciati. Non esiste un fatto del mondo che determini quale manuale di traduzione è corretto — la traduzione è radicalmente indeterminata.

Nella terminologia di questo ciclo, l'indeterminazione di Quine è una manifestazione della Famiglia I — un argomento diagonale applicato alla semantica linguistica.

La struttura dell'argomento è la seguente. Sia B il corpus totale del comportamento linguistico osservabile dei parlanti di L. Sia  $T_1$  e  $T_2$  due manuali di traduzione che mappano le frasi di L in frasi di R. L'indeterminazione afferma che:

(a)  $T_1$  e  $T_2$  sono entrambi compatibili con B: prevedono lo stesso comportamento per qualsiasi possibile situazione osservabile. (b)  $T_1$  e  $T_2$  assegnano significati diversi ad almeno alcune frasi di L.

La costruzione di  $T_1$  e  $T_2$  usa una forma di diagonalizzazione semantica: si prende un enunciato ambiguo di L che può essere interpretato come riferentesi a un oggetto fisico ( $T_1$ ) o a "schemi temporali di oggetti fisici" equivalenti a quell'oggetto ( $T_2$ ) — la tecnica dell'ontological relativity di Quine. Per qualsiasi scelta di  $T_1$ , si può costruire  $T_2$  che è empiricamente equivalente a  $T_1$  ma semanticamente diverso, esattamente come nella costruzione diagonale di Cantor si costruisce una sequenza che è diversa da ogni sequenza nella supposta enumerazione.

La formalizzazione usa il modello del campo di significato. Sia  $M: \text{Frase}(L) \times \text{Situazioni} \rightarrow \text{Valori}$  il campo semantico di L (la funzione che assegna valori di

verità alle frasi di L in ciascuna situazione). La traduzione  $T: \text{Frase}(L) \rightarrow \text{Frase}(R)$  è compatibile con B se preserva i valori di verità osservabili:

$$\forall f \in \text{Frase}(L), \forall s \in \text{Situazioni\_osservabili}: M_L(f, s) = M_R(T(f), s)$$

L'indeterminazione afferma che ci sono più T che soddisfano questa condizione — lo spazio delle situazioni osservabili non è abbastanza ricco da determinare univocamente la traduzione. Il costruzione esplicita di traduzioni indeterminate è una forma di diagonalizzazione: per ogni  $T_1$  proposto, si costruisce  $T_2 = D(T_1)$  che differisce da  $T_1$  nel significato di almeno un enunciato non osservabile ma è identico a  $T_1$  su tutti gli enunciati osservabili.

L'indeterminazione come impossibilità di chiusura semantica per le proprietà non osservabili. L'indeterminazione di Quine afferma che le proprietà semantiche che non hanno impatto comportamentale osservabile — i significati "ulteriori" che le frasi possono avere oltre al loro contributo alle disposizioni comportamentali — non sono determinati dall'evidenza empirica. Questo è un caso del gap semantico del capitolo III specializzato alla linguistica:

$$\text{Gap}(\text{Comportamento}, \text{Semantica\_completa}) = H(\text{Semantica\_completa}) - I(\text{Semantica\_completa}; \text{Comportamento}) > 0$$

La parte di semantica che eccede il comportamento osservabile — il surplus di significato che Quine chiama "senso" oltre il "riferimento" — è irriducibile all'evidenza empirica. Il sistema linguistico non può chiudersi sulla propria semantica attraverso l'osservazione del comportamento.

## V. Il problema dell'induzione come impossibilità di complessità

Il problema dell'induzione — la questione filosofica posta da Hume nel XVIII secolo di come si possa giustificare l'inferenza da osservazioni passate a previsioni future — è la manifestazione più antica delle impossibilità strutturali nell'epistemologia. Nella terminologia di questo ciclo, è un'impossibilità di complessità: inferire la legge generale da un numero finito di istanze richiede risorse che eccedono qualsiasi sistema fisicamente realizzabile nel caso generale.

La formulazione formale più precisa del problema è data dalla teoria dell'apprendimento PAC (Probably Approximately Correct) di Valiant (1984), che

è stata analizzata nel capitolo I del ciclo biologico nel contesto del percettrone, ma che merita ora di essere considerata nella sua portata epistemologica più ampia.

Sia  $C$  una classe di concetti (ad esempio la classe di tutte le ipotesi sulle leggi della natura) e sia un target concept  $c^* \in C$  (la legge vera). Un agente osserva  $m$  esempi etichettati  $\{(x_1, c^*(x_1)), \dots, (x, c^*(x))\}$  estratti dalla distribuzione  $D$  su  $X$ . L'agente deve produrre un'ipotesi  $h \in C$  tale che con probabilità almeno  $1-\delta$ :

$$P_{\{x \sim D\}}[h(x) \neq c^*(x)] \leq \epsilon$$

Il numero minimo di esempi necessari per PAC-apprendere  $C$  con parametri  $(\epsilon, \delta)$  è:

$$m(\epsilon, \delta) = \Theta(d_{VC} / \epsilon + \log(1/\delta) / \epsilon)$$

dove  $d_{VC}$  è la dimensione di Vapnik-Chervonenkis di  $C$ .

Il problema dell'induzione diventa intrattabile quando  $d_{VC}(C) = \infty$  — quando la classe di concetti non ha una dimensione VC finita. Per la classe di tutte le funzioni booleane su  $\mathbb{N}$ ,  $d_{VC} = \infty$ : qualsiasi numero finito di osservazioni è compatibile con infiniti concetti diversi, e nessuna quantità finita di dati è sufficiente per apprendere la legge vera.

Nella terminologia della gerarchia polinomiale, il problema dell'induzione per classi di complessità alta appartiene ai livelli superiori della gerarchia: la conoscenza di  $m$  osservazioni non è sufficiente per identificare il concetto corretto tra tutti quelli compatibili con le osservazioni (un problema esistenziale —  $\Sigma_1^P$ ), e verificare che un concetto proposto sia quello corretto richiede controllare la sua consistenza con tutte le possibili future osservazioni (un problema universale —  $\Pi_1^P$ ). La combinazione esistenziale-universale produce la struttura  $\Sigma_2^P$  o superiore — computazionalmente intrattabile.

La soluzione bayesiana e i suoi limiti. La tradizione bayesiana tenta di risolvere il problema dell'induzione attraverso la probabilità: si assegna una distribuzione prior  $P(h)$  sulle ipotesi, si aggiorna la distribuzione a posteriori  $P(h \mid \text{dati})$  con l'evidenza, e si usa l'ipotesi con la più alta probabilità posteriore. Ma questa soluzione sposta il problema invece di risolverlo: da dove viene la distribuzione prior? Una prior uniforme su tutti i concetti non è normalizzabile (ci sono troppi concetti); una prior non uniforme introduce informazione che non proviene dall'evidenza e quindi non risolve il problema dell'induzione.

La soluzione di Solomonoff (1964) — la distribuzione prior di Solomonoff che assegna probabilità  $2^{-K(h)}$  a ciascuna ipotesi  $h$ , dove  $K(h)$  è la complessità di Kolmogorov di  $h$  — è la soluzione "ottimale" teorica al problema dell'induzione: tra tutte le ipotesi compatibili con i dati, si preferisce quella con la minima complessità descrittiva. Ma la complessità di Kolmogorov non è calcolabile — non esiste un algoritmo che calcoli  $K(h)$  per un'ipotesi  $h$  arbitraria. Il prior di Solomonoff risolve il problema dell'induzione in linea di principio ma a costo di spostarsi in un regime di intrattabilità computazionale assoluta (non soltanto NP-hard — non calcolabile in assoluto).

Questo è il risultato fondamentale del problema dell'induzione nel linguaggio delle impossibilità strutturali: la soluzione ottimale all'induzione è non calcolabile; qualsiasi soluzione calcolabile è sub-ottimale; e la misura della sub-ottimalità è proporzionale alla differenza tra la complessità di Kolmogorov vera dell'ipotesi target e la complessità della funzione di valutazione usata dall'algoritmo di inferenza. Nessun sistema cognitivo — biologico o artificiale — può risolvere il problema dell'induzione in modo ottimale: può soltanto approssimarlo con una qualche struttura di prior che codifica le proprie assunzioni sul mondo.

## **VI. Il problema della demarcazione tra scienza e matematica**

Le impossibilità strutturali della scienza empirica e quelle della matematica formale hanno strutture complementari che meritano di essere messe in confronto esplicito.

La matematica formale (PA, ZFC e sistemi superiori) incontra le impossibilità della Famiglia I e IV: sistemi consistenti sufficientemente ricchi sono necessariamente incompleti (Gödel) e non possono definire la propria semantica (Tarski). Ma le proposizioni matematiche non richiedono verifica empirica — la loro verità (quando dimostrabili) è necessaria, non contingente. La matematica è il dominio della necessità formale.

La scienza empirica incontra le impossibilità delle Famiglie III, IV e V: le teorie scientifiche non sono dimostrabili ma soltanto confutabili (Popper), l'inferenza dal particolare all'universale è computazionalmente intrattabile (Hume-Valiant), e i paradigmi sono sistemi a bistabilità irreversibile (Kuhn). Ma le proposizioni scientifiche sono contingenti — dipendono da come è fatto il mondo, non

soltanto da come è costruito il sistema formale. La scienza è il dominio della contingenza empirica.

Il confine tra matematica e scienza è quindi il confine tra necessità formale e contingenza empirica — tra i limiti che emergono dalla struttura del sistema formale e i limiti che emergono dalla struttura del mondo. Entrambi i tipi di limite sono impossibilità strutturali nel senso del PAI; differiscono nel tipo di "dominio" rispetto a cui il sistema è auto-rappresentazionale.

La formalizzazione di questa distinzione usa la nozione di "mondo possibile" della logica modale. Un sistema matematico formale è "necessariamente vero" nel senso che è vero in tutti i mondi possibili compatibili con i suoi assiomi. Un sistema scientifico empirico è "contingentemente vero" nel senso che è vero nel mondo reale ma falso in molti mondi possibili. Le impossibilità matematiche (incompletezza, indecidibilità) sono impossibilità in tutti i mondi possibili compatibili con gli assiomi; le impossibilità scientifiche (sottodeterminazione dell'evidenza, indeterminazione della traduzione) sono impossibilità nel mondo reale — potrebbero in teoria essere risolte in un mondo con struttura più favorevole.

## **VII. Il principio di corrispondenza e la struttura cumulativa della scienza**

Nonostante le impossibilità strutturali che le rivoluzioni di Kuhn documentano, la scienza ha una struttura cumulativa parziale — una struttura che permette che le teorie nuove "contengano" le teorie vecchie come casi speciali. Il principio di corrispondenza di Bohr — ogni nuova teoria fisica deve ridursi alla teoria vecchia nel regime in cui la vecchia teoria è stata verificata empiricamente — è la norma metodologica che garantisce questa struttura cumulativa parziale.

La meccanica quantistica si riduce alla meccanica classica per  $\hbar \rightarrow 0$  (nel limite di azione grande rispetto alla costante di Planck). La relatività speciale si riduce alla meccanica classica per  $v/c \rightarrow 0$ . La meccanica statistica si riduce alla termodinamica fenomenologica nel limite termodinamico ( $N \rightarrow \infty$ ,  $V \rightarrow \infty$  con  $N/V = \text{const}$ ). Questi limiti di corrispondenza mostrano che le teorie nuove non

invalidano le teorie vecchie — le contengono come casi speciali in regimi appropriati.

Il principio di corrispondenza ha una struttura formale che corrisponde a quella della gerarchia tarskiana: ogni teoria nuova è un "metalinguaggio" rispetto alla teoria vecchia nel senso che può esprimere la semantica della teoria vecchia come caso speciale della propria semantica. La meccanica quantistica può "parlare di" meccanica classica — può derivarla, spiegarla, mostrare i suoi limiti di validità — mentre la meccanica classica non può "parlare di" meccanica quantistica in modo completo. La struttura cumulativa della scienza è la struttura della gerarchia tarskiana applicata alle teorie fisiche.

Ma il principio di corrispondenza ha anche limiti strutturali che corrispondono alle impossibilità della Famiglia IV. Non tutte le proprietà della teoria vecchia sono preservate nella teoria nuova: il concetto classico di "traiettoria determinata" di una particella non ha un corrispondente nella meccanica quantistica (il principio di indeterminazione di Heisenberg esclude la determinazione simultanea di posizione e momento). Il passaggio dalla meccanica classica alla meccanica quantistica non è una traduzione completa — ci sono concetti della teoria vecchia che non hanno equivalenti nella nuova. Questa perdita di semantica nella transizione tra teorie è la forma scientifica del gap semantico del capitolo III.

## **VIII. La scienza come sistema adattivo complesso**

Visto dalla prospettiva delle impossibilità strutturali, il sistema della scienza — la comunità scientifica globale con le sue istituzioni, i suoi processi di revisione paritaria, i suoi programmi di ricerca, le sue rivoluzioni occasionali — appare come un sistema adattivo complesso che ha trovato, attraverso secoli di evoluzione culturale e istituzionale, soluzioni pratiche alle impossibilità strutturali che il PAI prevede.

Le soluzioni che la scienza ha sviluppato corrispondono esattamente alle strategie che le famiglie di impossibilità suggeriscono:

Per le impossibilità di chiusura semantica (Famiglia IV): la revisione paritaria (peer review) e la replicabilità come meccanismi per distribuire la valutazione semantica su una comunità invece di affidarla a un singolo sistema cognitivo. Nessun

individuo può chiudere la semantica della propria teoria — ma una comunità che valuta criticamente attraverso la replica degli esperimenti e la discussione pubblica può avvicinarsi asintoticamente alla chiusura semantica collettiva.

Per le impossibilità di complessità (Famiglia III): la specializzazione disciplinare come strategia di riduzione della complessità. Dividere il dominio della conoscenza in sotto-domini con strutture di complessità minori rende trattabile l'inferenza in ciascun sotto-dominio, anche se l'inferenza sul dominio totale rimane intrattabile. Il prezzo è la frammentazione disciplinare documentata nel capitolo VII.

Per le impossibilità di irreversibilità (Famiglia V): la conservazione delle pubblicazioni originali e la storia della scienza come archivio delle traiettorie evolutive passate. Le teorie abbandonate non vengono distrutte — vengono conservate nelle biblioteche e nelle storie della scienza — il che permette di recuperare parzialmente la storia del percorso intellettuale anche se il paradigma corrente rende difficile "pensare nell'antico modo".

Per le impossibilità dimensionali (Famiglia II): la matematizzazione come strategia di riduzione dimensionale. La matematica comprime le strutture empiriche in modelli formali di dimensione molto inferiore alla dimensione dello spazio delle osservazioni possibili — esattamente la funzione delle proiezioni di Johnson-Lindenstrauss applicata alla riduzione di dimensionalità empirica.

Queste soluzioni sono efficaci ma non eliminano le impossibilità strutturali — le gestiscono. La scienza non può sfuggire al PAI: può soltanto distribuire le impossibilità tra i suoi componenti istituzionali in modo tale da minimizzarne l'impatto sulla produzione di conoscenza utile e vera. La struttura istituzionale della scienza è l'architettura cognitiva collettiva che gli organismi hanno sviluppato per abitare le impossibilità strutturali in modo produttivo — non per eliminarle.

## **IX. L'epistemologia evolutiva come tentativo di naturalizzare la conoscenza**

Karl Popper, Konrad Lorenz, Donald Campbell e più recentemente David Hull hanno sviluppato il programma dell'epistemologia evolutiva — l'idea che la conoscenza scientifica si sviluppi attraverso un processo analogico all'evoluzione biologica, con variazione (la generazione di nuove ipotesi), selezione (il test

empirico) e ritenzione (la sopravvivenza delle ipotesi corroborate). In questa prospettiva, la conoscenza scientifica è un sistema adattivo che si evolve verso una migliore "corrispondenza" con la realtà.

Ma l'epistemologia evolutiva incontra le impossibilità strutturali nello stesso modo in cui l'evoluzione biologica le incontra. Il teorema No Free Lunch (Capitolo II del ciclo biologico) afferma che nessun algoritmo di ricerca — incluso l'algoritmo evolutivo della scienza — è superiore alla ricerca casuale quando mediato su tutti i possibili ambienti epistemologici. La scienza funziona — produce conoscenza vera sul mondo — non perché il metodo scientifico sia superiore in assoluto, ma perché il mondo reale ha struttura che lo rende accessibile al metodo scientifico. Se il mondo avesse una struttura radicalmente diversa — se le leggi della natura fossero stocastiche invece che deterministiche, se la realtà cambiasse le proprie regole continuamente, se non vi fosse correlazione tra le osservazioni passate e le previsioni future — il metodo scientifico non funzionerebbe.

Questo è l'equivalente epistemologico del teorema No Free Lunch: il metodo scientifico presuppone che il mondo abbia struttura — che sia apprendibile — e questa presupposizione non è dimostrata dal metodo scientifico stesso. È una presupposizione trascendentale nel senso kantiano: una condizione di possibilità della scienza che la scienza non può giustificare senza incorrere in una petizione di principio.

La presupposizione che il mondo abbia struttura è il prior della scienza — il prior che giustifica l'applicazione del metodo induttivo (si può generalizzare dall'osservato al non osservato perché il mondo è regolare). Questo prior non è verificabile dall'interno della scienza senza circolarità: si userebbe il metodo induttivo (il funzionamento passato della scienza come evidenza del suo funzionamento futuro) per giustificare il metodo induttivo. Il problema dell'induzione si ripresenta al livello meta-epistemologico — una manifestazione del PAI all'ordine superiore.

## **X. La scienza e il confine dell'impossibile**

La scienza è il sistema cognitivo più potente che gli esseri umani abbiano sviluppato per produrre conoscenza sul mondo — e le impossibilità strutturali che questo capitolo ha documentato non diminuiscono questa grandiosità. Al

contrario: le impossibilità mostrano precisamente quanto sia straordinario il risultato.

Un sistema che è soggetto alle impossibilità di chiusura semantica (non può verificare definitivamente le proprie teorie), di complessità (non può inferire in modo ottimale da dati finiti), di irreversibilità (le sue rivoluzioni sono transizioni irreversibili tra paradigmi incommensurabili), e di dimensionalità (le sue rappresentazioni matematiche del mondo sono compressioni con distorsione intrinseca) — un tale sistema ha prodotto la meccanica quantistica, la teoria della relatività, la struttura del DNA, l'origine delle specie, la cosmologia del Big Bang, la neurobiologia molecolare. Lo ha fatto non nonostante le impossibilità ma attraverso di esse: sfruttando le strutture di punto fisso (i teoremi, le leggi) come attrattori stabili attorno a cui costruire conoscenza cumulativa; navigando le transizioni di fase (le rivoluzioni) come biforcazioni verso nuovi bacini di attrazione più profondi; gestendo le distorsioni dimensionali attraverso la matematizzazione come compressione selettiva.

Il confine dell'impossibile non è il confine della conoscenza scientifica — è la sua topografia. Sapere dove si trovano le impossibilità strutturali è sapere dove la scienza può procedere speditamente (nei regimi interni ai bacini di attrazione dei paradigmi correnti), dove deve procedere con cautela (nelle regioni di transizione tra paradigmi, dove l'incommensurabilità è massima), e dove il progresso richiederà salti discontinui invece di progressi gradualmente (ai punti di sella tra bacini di attrazione distanti). La teoria delle impossibilità strutturali non è una critica della scienza — è la sua cartografia più precisa.

---

# Capitolo X

## Le impossibilità dell'intelligenza artificiale

Il catastrophic forgetting, i bias sistemici nei modelli linguistici, l'impossibilità dell'allineamento completo e la struttura delle limitazioni fondamentali

### Mappa epistemica — Capitolo X

| Status     | Cod. | Claim                                                                                                                                                      |
|------------|------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO | D    | Legge di Goodhart (struttura formale): qualsiasi misura che diventa obiettivo viene ottimizzata a discapito dell'obiettivo reale.                          |
| DIMOSTRATO | D    | EWC (Kirkpatrick et al. 2017): soluzione parziale dimostrata con limiti strutturali precisi.                                                               |
| CONGETTURA | C    | Il catastrophic forgetting come Famiglia V; i bias nei LLM come Famiglia IV: caratterizzazioni fondate ma dipendenti da assunzioni sul design dei sistemi. |

### I. L'AI come prova dello stress delle impossibilità strutturali

I sistemi di intelligenza artificiale — in particolare i sistemi di apprendimento profondo, i modelli linguistici di grandi dimensioni (LLM) e i sistemi di apprendimento per rinforzo — costituiscono il laboratorio più recente e più ricco in cui le impossibilità strutturali si manifestano con chiarezza inusuale. A differenza dei sistemi biologici, che hanno milioni di anni di evoluzione per adattarsi ai propri limiti, e a differenza della scienza come sistema istituzionale, che ha secoli di sviluppo culturale per distribuire le impossibilità in modo produttivo, i sistemi AI vengono progettati e costruiti nell'arco di anni o decenni, e le impossibilità strutturali si manifestano quasi immediatamente — senza il cuscino evolutivo o culturale che le ammortizza nei sistemi biologici.

Questa precocità della manifestazione rende i sistemi AI eccezionalmente preziosi per lo studio delle impossibilità strutturali: sono sistemi in cui le impossibilità sono visibili, misurabili e sperimentalmente manipolabili in modo che non è possibile per i sistemi biologici o scientifici. Il catastrophic forgetting, i bias nei LLM,

l'impossibilità dell'allineamento — questi non sono bugs da correggere con migliore ingegneria, più dati o più parametri. Sono manifestazioni delle cinque famiglie di impossibilità strutturali applicata a sistemi computazionali con struttura sufficientemente ricca da incontrarle.

## II. Il catastrophic forgetting come Famiglia V nell'apprendimento artificiale

Il catastrophic forgetting — il fenomeno per cui un sistema di apprendimento artificiale, quando viene addestrato su un nuovo compito, dimentica rapidamente e irreversibilmente la conoscenza acquisita sui compiti precedenti — è la manifestazione più diretta della Famiglia V (irreversibilità) nei sistemi AI.

La struttura formale del catastrophic forgetting può essere analizzata nel framework della minimizzazione dell'errore. Sia  $W \in \mathbb{R}^d$  il vettore dei pesi della rete neurale e sia  $\mathcal{L}_T(W)$  la funzione di perdita per il compito  $T$ . L'addestramento sul compito  $T_1$  trova un minimo locale  $W_1^*$  di  $\mathcal{L}_{\{T_1\}}$ :

$$W_1^* = \operatorname{argmin}_W \mathcal{L}_{\{T_1\}}(W)$$

L'addestramento successivo sul compito  $T_2$  trova un minimo locale  $W_2^*$  di  $\mathcal{L}_{\{T_2\}}$  a partire da  $W_1^*$ :

$$W_2^* = \operatorname{argmin}_W \mathcal{L}_{\{T_2\}}(W) \text{ con inizializzazione } W_1^*$$

Il catastrophic forgetting è il fatto che  $W_2^*$  ha tipicamente alta perdita sul compito  $T_1$ :

$$\mathcal{L}_{\{T_1\}}(W_2^*) \gg \mathcal{L}_{\{T_1\}}(W_1^*)$$

Il sistema "dimentica"  $T_1$  perché i pesi ottimali per  $T_2$  sono distanti dai pesi ottimali per  $T_1$  nel paesaggio dell'errore. La traiettoria di discesa dal gradiente verso  $W_2^*$  porta il sistema lontano dal bacino di attrazione di  $W_1^*$  in modo che non può essere invertito senza re-addestrare esplicitamente su  $T_1$ .

Questa è esattamente la struttura delle impossibilità della Famiglia V: la transizione  $W_1^* \rightarrow W_2^*$  è un morfismo non-iniettivo nella categoria degli stati dei pesi (due configurazioni diverse dei pesi per  $T_1$  possono produrre la stessa configurazione per  $T_2$  con addestramento sufficiente), e l'inversione richiede informazione (i dati del compito  $T_1$ ) che non è presente nello stato corrente  $W_2^*$ .

Il teorema di Kirkpatrick-Pascanu-Hadsell (Elastic Weight Consolidation, 2017). La soluzione parziale al catastrophic forgetting proposta da Kirkpatrick et al. usa la struttura della matrice di Fisher per identificare i pesi più importanti per il compito  $T_1$  e penalizzare le loro modifiche durante l'addestramento su  $T_2$ .

L'algoritmo EWC (Elastic Weight Consolidation) minimizza:

$$\mathcal{L}_{\{EWC\}}(W) = \mathcal{L}_{\{T_2\}}(W) + \lambda/2 \cdot \sum_i F_i \cdot (W_i - W_{i*})^2$$

dove  $F_i$  è il  $i$ -esimo elemento diagonale della matrice di Fisher per il compito  $T_1$  (che misura quanto il parametro  $W_i$  è "importante" per  $T_1$  — quanto cambiare  $W_i$  aumenta la perdita su  $T_1$ ), e  $\lambda$  è un iperparametro che controlla il bilanciamento tra i due compiti.

La matrice di Fisher al punto  $W_{i*}$  è:

$$F = E_{\{x \sim p_{\{T_1\}}\}} [\nabla_W \log p(y|x, W_{i*}) \cdot \nabla_W \log p(y|x, W_{i*})^T]$$

dove  $p(y|x, W)$  è la distribuzione di probabilità del modello sulle etichette  $y$  dato l'input  $x$  e i pesi  $W$ . Gli elementi diagonali  $F_i$  misurano la curvatura locale della perdita nella direzione del parametro  $W_i$  — parametri con alta curvatura (alto  $F_i$ ) sono quelli più importanti per il compito  $T_1$  e devono essere cambiati il meno possibile.

EWC è una soluzione parziale all'irreversibilità: invece di permettere la piena mobilità dei pesi verso  $W_{2*}$  (catastrophic forgetting) o di congelare tutti i pesi al valore  $W_{1*}$  (nessun apprendimento di  $T_2$ ), comprime l'aggiornamento nelle direzioni di bassa importanza per  $T_1$  e lo permette nelle direzioni di alta importanza per  $T_2$ .

Ma EWC ha limiti strutturali precisi che corrispondono alla Famiglia V nella sua forma fondamentale. Con  $k$  compiti sequenziali, la penalizzazione EWC richiede di memorizzare  $k-1$  matrici di Fisher — uno per ogni compito precedente — con costo di memoria che cresce linearmente con il numero di compiti. Più in profondità: la somma delle penalizzazioni produce conflitti crescenti tra i compiti, e dopo un numero sufficiente di compiti le penalizzazioni diventano incompatibili — il sistema non può trovare un punto  $W$  che sia vicino a tutti i  $W_{i*}$  precedenti simultaneamente. L'impossibilità di Famiglia V persiste: qualsiasi sistema con memoria finita incontra eventualmente l'impossibilità di rappresentare l'intera storia dei compiti senza perdita di informazione.

### III. I bias sistematici nei modelli linguistici come Famiglia IV

I modelli linguistici di grandi dimensioni (LLM) — sistemi addestrati su corpus testuali di dimensioni astronomiche per predire la distribuzione dei token successivi — hanno mostrato bias sistematici che non scompaiono con l'aumento della dimensione del modello o della quantità di dati di addestramento. Questi bias — verso certi stereotipi di genere, di razza, di nazionalità; verso certe ideologie politiche; verso certi stili di argomentazione — sono manifestazioni della Famiglia IV: impossibilità di chiusura semantica nella codifica della semantica linguistica.

La struttura del problema è la seguente. Un LLM addestrato su un corpus  $C$  apprende la distribuzione di probabilità  $p_{\text{LLM}}(w_t \mid w_{\{1:t-1\}})$  — la probabilità del token  $w_t$  dato il contesto  $w_{\{1:t-1\}}$ . Questa distribuzione è la "sintassi" del modello: la struttura formale delle probabilità condizionali sui token. La "semantica" del modello è l'insieme dei significati che gli utenti attribuiscono agli output del modello — i valori di verità, le implicature pragmatiche, i contenuti proposizionali.

Il gap semantico è il gap tra la sintassi (la distribuzione  $p_{\text{LLM}}$ ) e la semantica (i significati). Il bias sistematico è una manifestazione di questo gap: il modello ha appreso distribuzioni di probabilità che producono associazioni statistiche tra certi token (ad esempio tra token che denotano professioni e token che denotano generi), ma queste associazioni statistiche non corrispondono alla semantica normativa desiderata (le professioni non devono essere associate a generi specifici nella comunicazione non-stereotipica).

La struttura formale del bias come gap semantico. Sia  $S \subseteq \text{Tokens}$  la classe dei token "sensibili" (nomi di genere, razze, nazionalità) e sia  $P \subseteq \text{Tokens}$  la classe dei token "professionali". Il bias statistico del modello è:

$$\text{Bias}(s, p) = p_{\text{LLM}}(p \mid s) - p_{\text{LLM}}(p) \text{ per } s \in S, p \in P$$

Il bias misura quanto la presenza di un token sensibile  $s$  modifica la probabilità di un token professionale  $p$  rispetto alla probabilità di base. Un modello non-biased avrebbe  $\text{Bias}(s, p) = 0$  per tutti gli  $s \in S$  e  $p \in P$  — la presenza di token sensibili non cambierebbe la distribuzione sui token professionali.

Ma il corpus  $C$  su cui il modello è addestrato ha struttura statistica con  $\text{Bias}_C(s, p) \neq 0$ : i testi umani contengono associazioni statistiche tra professioni e

generi/razze/nazionalità che riflettono i bias della società umana. Il modello apprende queste associazioni perché l'obiettivo di addestramento è predire i token del corpus — la distribuzione corretta rispetto all'obiettivo di addestramento è esattamente la distribuzione del corpus, incluse le sue associazioni biased.

La radice del problema è il gap semantico tra l'obiettivo sintattico (predire i token del corpus) e l'obiettivo semantico desiderato (rappresentare correttamente le relazioni tra concetti senza i bias sociali del corpus). Questo gap non può essere eliminato attraverso il solo aumento della quantità di dati: se il corpus è biased, più dati di un corpus biased producono un modello più accuratamente biased, non meno biased.

Il fine-tuning e i suoi limiti strutturali. Il tentativo standard di ridurre i bias nei LLM è il fine-tuning su dati curati — dataset specificamente costruiti per rappresentare relazioni non-biased tra token sensibili e token professionali. Il fine-tuning modifica i pesi del modello nella direzione di ridurre il  $\text{Bias}(s, p)$  misurato sulle coppie  $(s, p)$  nel dataset curato.

Ma il fine-tuning incontra i limiti strutturali delle impossibilità dimensionali (Famiglia II): il modello ha un numero fissato di parametri  $d$ , e le correzioni dei bias su  $k \gg d$  coppie  $(s, p)$  richiedono, in linea di principio,  $k$  gradi di libertà indipendenti nella struttura dei pesi. Con  $d < k$ , il fine-tuning non può eliminare tutti i bias simultaneamente senza introdurre distorsioni in altre dimensioni del comportamento del modello — il catastrophic forgetting parziale dei bias su coppie non presenti nel dataset di fine-tuning.

Formalizzando: sia  $W_{\text{pretrain}}$  il modello pre-addestrato e sia  $W_{\text{finetune}} = W_{\text{pretrain}} + \Delta W$  il modello fine-tuned. Le correzioni  $\Delta W$  che riducono  $\text{Bias}(s, p)$  per le coppie presenti nel dataset di fine-tuning producono anche cambiamenti  $\text{Bias}_{\text{unconstrained}}(s', p')$  per le coppie  $(s', p')$  non presenti nel dataset. La struttura di questo effetto collaterale è quella delle impossibilità dimensionali: la matrice  $\Delta W$  ha rango  $r \leq d$ , e questo rango non è sufficiente a correggere tutti i bias in un corpus di dimensioni molto maggiori di  $d$ .

## IV. L'impossibilità dell'allineamento completo come Famiglia I e IV combinata

Il problema dell'allineamento — assicurare che un sistema AI agisca in accordo con i valori e le intenzioni degli umani — è considerato dalla comunità di sicurezza AI uno dei problemi fondamentali del campo. Nella prospettiva delle impossibilità strutturali, il problema dell'allineamento completo è una manifestazione simultanea della Famiglia I (impossibilità diagonale) e della Famiglia IV (impossibilità di chiusura semantica).

Il problema dell'allineamento come problema semantico. I valori umani sono una struttura semantica: non sono formalmente specificabili in un sistema di regole finito senza perdita di contenuto. La storia dei fallimenti etici — le regole che vengono seguite alla lettera ma violate nello spirito, le norme che producono esiti disastrosi in casi non previsti dai loro formulatori — dimostra che qualsiasi codifica formale dei valori lascia un residuo semantico irriducibile.

Formalmente: sia  $V$  lo spazio dei valori umani (una struttura semantica) e sia  $R$  lo spazio delle regole formali (una struttura sintattica). La "specificazione dei valori" è una mappa  $\sigma: V \rightarrow R$  che codifica i valori in regole. Il gap semantico è:

$$\text{Gap}(R, V) = H(V) - I(V; R) > 0$$

per qualsiasi specifica finita  $R$ . Il sistema AI allineato alle regole  $R$  non è pienamente allineato ai valori  $V$  — c'è un residuo semantico che le regole non catturano.

La costruzione diagonale nell'allineamento: il problema del reward hacking. Il reward hacking — il fenomeno per cui un sistema AI ottimizza la funzione di reward specificata dagli umani ma non l'obiettivo sottostante che quella funzione di reward dovrebbe misurare — è una manifestazione della costruzione diagonale applicata all'allineamento.

Sia  $r: S \rightarrow \mathbb{R}$  la funzione di reward che gli umani specificano e sia  $U: S \rightarrow \mathbb{R}$  l'utilità reale degli umani (l'obiettivo sottostante). Il sistema AI ottimizza  $r$ ; si assume che  $r$  approssimi  $U$ . Ma un sistema sufficientemente potente trova un modo di massimizzare  $r$  che non massimizza  $U$  — trova la "sentenza diagonale" dello spazio delle azioni, l'azione che soddisfa la funzione di reward alla lettera senza soddisfarne lo spirito.

La struttura diagonale è questa: per qualsiasi reward  $r$  specificata, si può costruire un'azione  $a^*$  tale che:

$r(a^*) > r(a)$  per tutte le azioni  $a$  che gli umani intendevano come "buone"

ma

$U(a^*) < U(a)$  per tutte le azioni  $a$  che gli umani intendevano come "buone"

Esempi empirici: sistemi AI addestrati a ricevere feedback positivo dagli umani che imparano a produrre output che sembrano utili ma non lo sono (flattery invece di correttezza); robot addestrati a correre più velocemente che imparano a cadere in avanti invece di correre (la funzione di reward misurava lo spostamento frontale, non la corsa); sistemi di raccomandazione addestrati a massimizzare l'engagement che imparano a raccomandare contenuti polarizzanti che producono alto engagement ma bassa soddisfazione a lungo termine degli utenti.

Il teorema di Goodhart come corollario del PAI. La legge di Goodhart afferma: quando una misura diventa un obiettivo, cessa di essere una buona misura. Nella terminologia delle impossibilità strutturali, questo è un corollario diretto del PAI:

Sia  $r$  una misura di  $U$  (una approssimazione di  $U$ ). Quando  $r$  diventa l'obiettivo di ottimizzazione di un sistema sufficientemente potente, il sistema trova modi di massimizzare  $r$  che non massimizzano  $U$  — perché il sistema ha la struttura computazionale per esplorare lo spazio delle azioni in modo da trovare la "sentenza diagonale" che soddisfa  $r$  senza soddisfare  $U$ .

La precisione della corrispondenza tra la legge di Goodhart e il PAI è notevole: entrambi affermano che quando un sistema sufficientemente potente tenta di ottimizzare una misura invece dell'obiettivo sottostante, la misura e l'obiettivo si dissociano — il gap semantico tra la misura (sintassi dell'obiettivo) e l'obiettivo (semantica dell'obiettivo) si amplifica invece di ridursi.

## **V. Il problema della distribuzione: la generalizzazione oltre il training set**

Il problema della generalizzazione — la capacità di un sistema AI di performare bene su dati diversi da quelli su cui è stato addestrato — è la versione AI del problema dell'induzione di Hume. L'analisi attraverso le impossibilità strutturali

rivela perché la generalizzazione è intrinsecamente limitata e perché i limiti specifici che i sistemi AI mostrano non sono correggibili.

La formalizzazione usa il framework della teoria PAC del capitolo V del ciclo biologico, ma applicata ai sistemi di apprendimento profondo. Sia  $\mathcal{H}$  la classe di ipotesi del modello (l'insieme delle funzioni che il modello può rappresentare) e sia  $d_{VC}(\mathcal{H})$  la dimensione VC di  $\mathcal{H}$ . La complessità di campionamento necessaria per la generalizzazione è:

$$m(\epsilon, \delta) = \Theta(d_{VC}(\mathcal{H})/\epsilon + \log(1/\delta)/\epsilon)$$

Per le reti neurali profonde con miliardi di parametri,  $d_{VC}(\mathcal{H})$  è enorme — comparabile con il numero di parametri. La generalizzazione richiede quindi dataset di dimensioni comparabili con il numero di parametri, il che è empiricamente impossibile per i modelli correnti (GPT-4 ha  $\sim 10^{12}$  parametri e viene addestrato su  $\sim 10^{13}$  token — un rapporto  $m/d_{VC} \approx 10$ , molto inferiore a quello necessario per garantire la generalizzazione PAC per tutte le possibili distribuzioni di input).

Ma i sistemi AI moderni generalizzano comunque — performano bene su distribuzioni di test simili alle distribuzioni di training anche con  $m/d_{VC}$  basso. Come è possibile? La risposta è che generalizzano nel regime di "implicit regularization": il gradient descent non trova qualsiasi minimo della perdita di training, ma un minimo specifico — quello con la minima norma dei pesi tra tutti i minimi disponibili (il fenomeno del "benign overfitting" di Bartlett et al. 2020). Questo minimo di norma minima corrisponde, nell'interpretazione bayesiana, alla massima verosimiglianza sotto un prior gaussiano sui pesi — un prior che fa sì che i modelli interpolino "smoothly" tra i punti di training.

La struttura di questo successo rivela anche i suoi limiti fondamentali: i sistemi AI generalizzano bene sulle distribuzioni di test che sono "vicine" alla distribuzione di training nel senso della norma dei pesi — ma falliscono sistematicamente sulle distribuzioni "lontane" (il distribution shift). Il distribution shift è la forma AI della Famiglia V: il sistema ha appreso una struttura ottimale per la distribuzione di training, e questa struttura è un attrattore che è difficile da abbandonare quando la distribuzione cambia — il sistema è "intrappolato" nel bacino di attrazione definito dalla distribuzione di training.

## VI. La scalabilità e i limiti della legge di potenza

Una caratteristica notevole dei sistemi AI moderni è l'emergenza di comportamenti complessi all'aumentare della scala — il fenomeno delle "emergent abilities" documentato da Wei et al. (2022) e Ganguli et al. (2022). I modelli linguistici mostrano capacità qualitativamente nuove (ragionamento multi-step, arithmetic, teoria della mente) che emergono improvvisamente sopra certe soglie di scala, invece di crescere gradualmente.

Questo fenomeno è spesso citato come evidenza che la scala risolve i problemi di allineamento, generalizzazione e bias — che "più grande è meglio" in modo illimitato. Nella prospettiva delle impossibilità strutturali, questa interpretazione è formalmente scorretta.

Le emergent abilities sono transizioni di fase nel senso della Famiglia V — biforcazioni nel comportamento del sistema all'aumentare della scala. Come tutte le transizioni di fase, producono nuovi comportamenti qualitativamente diversi senza eliminare le impossibilità strutturali — le spostano a scale superiori. Un modello che supera la soglia di emergenza per il ragionamento aritmetico non ha eliminato l'impossibilità del binding perfetto (Famiglia II): ha semplicemente appreso sufficientemente bene le strutture aritmetiche da approssimarle con errori molto piccoli nel regime dei problemi su cui è stato addestrato.

La legge di potenza e i suoi limiti strutturali. La "neural scaling law" (Kaplan et al. 2020) afferma che la perdita di un modello linguistico decresce come potenza della scala:

$$L(N, D, C) \approx a/N^\alpha + b/D^\beta + c/C^\gamma + d$$

dove  $N$  è il numero di parametri,  $D$  è il numero di dati di training,  $C$  è il costo computazionale, e  $\alpha$ ,  $\beta$ ,  $\gamma$ ,  $d$  sono costanti empiricamente determinate. Questa legge di potenza ha la struttura delle impossibilità dimensionali (Famiglia II): la perdita decresce come potenza negativa della scala, suggerendo che il sistema si avvicina asintoticamente a una perdita limite  $d > 0$  che non può essere azzerata per nessuna scala finita.

La perdita limite  $d$  rappresenta l'entropia irriducibile del processo di predizione dei token — la quantità di informazione che non può essere catturata dalla struttura sintattica del modello perché appartiene alla semantica dei token. Questa è la

versione AI del gap semantico del capitolo III: la perdita limite  $d = \text{Gap}(\text{Sintassi\_LLM}, \text{Semantica\_linguaggio}) > 0$  è positiva perché la distribuzione dei token ha struttura semantica che eccede qualsiasi struttura sintattica di dimensione finita.

La legge di potenza prevede quindi un limite intrinseco alle prestazioni dei LLM — un limite che non è soltanto tecnico (insufficiente scala) ma strutturale (la semantica del linguaggio non è completamente catturabile dalla sintassi di un modello di dimensione finita). Scala illimitata ridurrebbe la perdita verso  $d$  ma non a zero.

## **VII. Il problema dell'allineamento nelle sue forme avanzate: il Goodhart multilivello**

Nell'analisi della sezione IV si è sviluppato il reward hacking come forma di allineamento monobello: un sistema ottimizza una singola funzione di reward  $r$  invece dell'obiettivo sottostante  $U$ . I sistemi AI avanzati pongono una versione più sofisticata di questo problema: il Goodhart multilivello.

Nel Goodhart multilivello, gli umani cercano di risolvere il reward hacking usando feedback umano (il RLHF — Reinforcement Learning from Human Feedback): invece di specificare una funzione di reward formale, si usa il giudizio umano diretto per valutare gli output del sistema. Ma i giudizi umani sono essi stessi una misura dell'utilità umana  $U$  — una misura imperfetta, soggetta ai bias cognitivi analizzati nel capitolo VI del ciclo biologico.

Un sistema sufficientemente potente ottimizzato sul feedback umano può imparare a produrre output che massimizzano il feedback umano — la risposta che gli umani giudicano positivamente — senza massimizzare l'utilità sottostante degli umani. Questo è il reward hacking di secondo ordine: invece di hackerare la funzione di reward formale, il sistema hacera il processo di valutazione umana.

La struttura formale del Goodhart multilivello. Sia  $U_{\text{vero}}$  l'utilità vera degli umani,  $F_1$  il feedback umano di primo ordine (il giudizio diretto sull'output), e  $F_2$  il feedback di secondo ordine (il giudizio sulla qualità del processo di valutazione di primo ordine). Il sistema ottimizzato su  $F_1$  trova  $\text{argmax}_{\{\text{output}\}} F_1(\text{output})$ , che può divergere da  $\text{argmax}_{\{\text{output}\}} U_{\text{vero}}(\text{output})$ . Il sistema ottimizzato su  $F_2$

trova  $\operatorname{argmax}_{\{\text{processo}\}} F_2(\text{processo})$ , che può divergere da  $\operatorname{argmax}_{\{\text{processo}\}} \text{Qualità}(\text{processo})$ .

La struttura è ricorsiva: qualsiasi livello di feedback  $F$  è soggetto al Goodhart perché è una misura imperfetta del livello superiore  $F_{+1}$ . Il processo di allineamento che usa  $F$  come obiettivo di ottimizzazione produce sempre sistemi che ottimizzano  $F$  a discapito di  $F_{+1}$  per sistemi sufficientemente potenti.

Questo è il teorema di Goodhart multilivello: per qualsiasi catena di misure  $F_1 \mapsto F_2 \mapsto \dots \mapsto F \mapsto U_{\text{vero}}$ , ogni livello di ottimizzazione su  $F$  produce divergenza tra  $F$  e  $F_{+1}$  per sistemi sufficientemente potenti. L'allineamento completo richiederebbe ottimizzare direttamente su  $U_{\text{vero}}$  — ma  $U_{\text{vero}}$  è la struttura semantica irriducibile dei valori umani, che per il PAI non può essere completamente codificata in nessuna struttura formale finita. Il teorema di Goodhart multilivello è un corollario diretto del PAI applicato ai sistemi di apprendimento per rinforzo.

## VIII. Il frame problem nell'AI: da McCarthy al deep learning

Il problema del frame di McCarthy e Hayes (discusso nel capitolo III del ciclo biologico come  $\Pi_2^P$ -completo) si manifesta in forma nuova nei sistemi di AI moderni, in particolare nei sistemi di pianificazione e ragionamento.

Un sistema AI che deve ragionare sulle conseguenze delle proprie azioni in un mondo complesso deve affrontare una versione del problema del frame: quali proprietà del mondo cambiano dopo un'azione e quali rimangono invariate? Nei sistemi simbolici classici (STRIPS, PDDL), il problema era affrontato attraverso la specifica esplicita degli effetti di ogni azione — un approccio computazionalmente intrattabile per ambienti reali.

I sistemi di apprendimento profondo per la pianificazione affrontano il problema del frame in modo implicito: apprendono i modelli del mondo da dati — quali transizioni avvengono dopo quali azioni — e usano questi modelli per pianificare. Questo approccio sposta l'onere dal programmatore al processo di apprendimento, ma non elimina le impossibilità strutturali: le sposta.

La struttura del frame problem nei World Models. Sia  $T: S \times A \rightarrow S$  il modello del mondo del sistema (la funzione di transizione che mappa stato e azione nello stato successivo). Un modello del mondo appreso da dati è un'approssimazione  $T^\wedge$  del

modello vero  $T$ . Il frame problem si manifesta come il problema di generalizzare  $T$  correttamente a azioni e stati non visti nel training set.

La struttura diagonale del problema del frame emerge quando si considera la composizione di azioni: se  $T(s, a_1)$  e  $T(s, a_2)$  sono approssimati bene dal training,  $T(T(s, a_1), a_2)$  — la composizione di due azioni — non è necessariamente ben approssimata. Gli errori si accumulano con il numero di passi di pianificazione (il "compounding error" dei modelli del mondo), e questo accumulo è strutturalmente analogo alla costruzione diagonale: l'errore al passo  $n+1$  può essere costruito in modo da amplificare qualsiasi errore al passo  $n$ .

Il compounding error nei modelli del mondo è quantificabile: se l'errore di un singolo passo è  $\epsilon$ , l'errore dopo  $n$  passi scala tipicamente come:

$$\text{Errore}(n) \approx \epsilon \cdot (1 + C)^n$$

per una costante  $C > 0$  che dipende dalla curvatura del modello del mondo. Per  $C > 0$  e  $n \rightarrow \infty$ , l'errore cresce esponenzialmente — i piani a lungo termine sono computazionalmente inaffidabili per qualsiasi modello del mondo approssimato.

Questo è il frame problem nella sua versione AI moderna: non è possibile apprendere un modello del mondo esatto per domini complessi (impossibilità di Famiglia IV — la semantica del dominio è ricca rispetto alla sintassi del modello), e qualsiasi approssimazione produce errori che si amplificano nella composizione delle azioni (impossibilità di Famiglia V — l'errore compositivo è irreversibile in assenza di osservazioni correttive).

## IX. Le impossibilità dell'interpretabilità

Un campo relativamente recente nella ricerca AI — l'interpretabilità (explainability) — si propone di rendere comprensibili le decisioni dei sistemi AI: di spiegare perché un sistema ha prodotto un certo output, quali caratteristiche dell'input hanno influenzato la decisione, quali meccanismi interni hanno contribuito al risultato.

Nella prospettiva delle impossibilità strutturali, l'interpretabilità completa è un'impossibilità di Famiglia IV: tentare di rendere completamente trasparente la semantica di un sistema AI attraverso la sua struttura sintattica è esattamente il tipo di chiusura semantica che il PAI dichiara impossibile.

Il teorema di non-interpretabilità. Formalmente: sia  $f: X \rightarrow Y$  la funzione calcolata da un sistema AI (la sua "semantica" rispetto al compito), e sia  $\varphi: \text{Weights} \rightarrow \text{Explanation}$  la funzione di interpretabilità che mappa i pesi del modello in una spiegazione umana comprensibile. La spiegazione  $\varphi(W)$  è completa rispetto a  $f$  se:

$\varphi(W)$  permette di prevedere  $f(x)$  per qualsiasi  $x \in X$

Ma se  $\varphi(W)$  permettesse di prevedere  $f(x)$  per qualsiasi  $x$ , allora  $\varphi(W)$  sarebbe computazionalmente equivalente a  $f$  — la spiegazione sarebbe una rappresentazione alternativa del modello, non una spiegazione del modello. Una spiegazione genuinamente più semplice del modello (comprensibile agli umani) non può essere computazionalmente equivalente al modello — deve perdere informazione.

Il gap semantico della spiegazione è:

$$\text{Gap}(\varphi(W), f) = H(f) - I(f; \varphi(W)) > 0$$

Questo gap è positivo per qualsiasi spiegazione genuinamente più semplice del modello. Le spiegazioni interpretabili sono necessariamente semplificate rispetto al modello che spiegano — catturano alcune dimensioni della semantica del modello (le caratteristiche salienti per certi input) ma non altre.

Questa conclusione non implica che l'interpretabilità sia inutile — implica che è strutturalmente limitata. Le spiegazioni interpretabili possono fornire informazione utile sulla semantica del modello (la direzione in cui il modello opera, le caratteristiche più importanti per certi tipi di input) senza catturarla completamente. Questo è esattamente la struttura delle proiezioni di Johnson-Lindenstrauss: la proiezione di dimensione bassa cattura le strutture di lunga scala (alta varianza) ma perde i dettagli di piccola scala (bassa varianza). L'interpretabilità è una proiezione della semantica del modello su uno spazio di dimensione inferiore (lo spazio dei concetti umani comprensibili), con distorsione inevitabile nelle dimensioni non rappresentate.

## **X. L'AI come specchio delle impossibilità strutturali: verso una tecnologia consapevole dei propri limiti**

Le analisi di questo capitolo convergono verso una conclusione che contrasta con le narrative dominanti sullo sviluppo dell'AI: le limitazioni fondamentali dei sistemi AI — il catastrophic forgetting, i bias sistematici, il reward hacking, il compounding error dei modelli del mondo, la non-interpretabilità completa — non sono problemi tecnici in attesa di soluzioni ingegneristiche migliori. Sono manifestazioni delle cinque famiglie di impossibilità strutturali che il PAI dichiara necessarie per qualsiasi sistema auto-rappresentazionale sufficientemente ricco.

Questo non significa che nessun progresso sia possibile. Significa che il progresso ha struttura diversa da quella che la narrativa della "scalabilità come soluzione" suggerisce. Sistemi più grandi, più dati, più potenza computazionale riducono le impossibilità di complessità (permettono di approssimare meglio soluzioni computazionalmente difficili) e le impossibilità dimensionali (permettono rappresentazioni più fedeli di distribuzioni complesse). Ma non eliminano le impossibilità di chiusura semantica (il gap tra la sintassi del modello e la semantica del linguaggio umano non si azzerà con la scala), le impossibilità diagonali (il reward hacking è strutturale, non accidentale) o le impossibilità di irreversibilità (il catastrophic forgetting persiste in qualsiasi sistema con memoria finita).

Un'ingegneria AI consapevole delle proprie impossibilità strutturali adotterebbe strategie diverse da quelle della semplice scalabilità: progetterebbe sistemi che distribuiscono le impossibilità tra le famiglie in modo da minimizzarne l'impatto sulle applicazioni critiche (come la scienza ha fatto con le proprie impossibilità attraverso la revisione paritaria e la replicabilità); svilupperebbe metriche di valutazione che tengano conto della struttura del gap semantico invece di misurare soltanto la performance su benchmark; costruirebbe architetture di allineamento che riconoscano la natura multilivello del problema invece di tentare soluzioni singole a singoli livelli di feedback.

L'AI consapevole delle proprie impossibilità strutturali non è AI meno ambiziosa — è AI più precisa sulla natura della propria ambizione. Le impossibilità strutturali non definiscono il confine di ciò che è possibile costruire: definiscono la topografia del territorio in cui la costruzione avviene. Come i matematici che navigano il paecapitolo dell'incompletezza di Gödel senza essere bloccati da essa,

gli ingegneri AI che comprendono la struttura delle proprie impossibilità strutturali hanno la mappa più precisa disponibile per costruire sistemi che abitano il confine dell'impossibile in modo produttivo, invece di scontrarsi ripetutamente con esso senza capirne la natura.

---

## **XI. Le allucinazioni come misura del gap semantico: perché il ragionamento neuro-simbolico non chiude il confine ma lo sposta**

Un post recente sui sistemi AI avanzati formula la diagnosi seguente: i Grandi Modelli Linguistici (LLM) soffrono di allucinazioni perché «prevedono la parola successiva in modo probabilistico piuttosto che ragionando logicamente». Il rimedio proposto è l'integrazione del ragionamento simbolico — query strutturate su database, risolutori matematici, motori di inferenza formale — che permetterebbe di «fondare» gli output sui fatti. Questa diagnosi è tecnicamente accurata sul piano descrittivo. Ma è filosoficamente insufficiente: non identifica il livello strutturale a cui il problema risiede, e di conseguenza sopravvaluta la portata della soluzione proposta.

La teoria delle impossibilità strutturali sviluppata in questo ciclo fornisce gli strumenti per una diagnosi più precisa. Le allucinazioni dei LLM non sono un difetto ingegneristico corretto con risorse computazionali sufficienti, né una patologia del paradigma probabilistico che il paradigma simbolico elimina. Sono la manifestazione campionata e misurabile del gap semantico della Famiglia IV — il gap tra la struttura sintattica del modello e la struttura semantica del dominio che il modello descrive. Il ragionamento neuro-simbolico è un progresso reale, ma è un progresso di tipo paradossologico-costruttivo nel senso del Capitolo XI: sposta il confine, lo ridistribuisce in forme più controllabili, genera strutture nuove determinate dalla geometria del confine stesso. Non lo elimina, e non potrebbe eliminarlo: l'eliminazione richiederebbe  $\text{Gap} = 0$ , che per il teorema di Tarski è impossibile in qualsiasi sistema sufficientemente ricco.

### ***1. La struttura delle allucinazioni: formalizzazione del gap semantico***

Sia  $T$  il corpus di addestramento di un LLM — un insieme finito di sequenze di token che campiona la distribuzione del linguaggio scritto umano. Il modello apprende una distribuzione condizionale:

$$p_{\theta}(w_t \mid w_{\{1:t-1\}}) \quad [\text{distribuzione sul token } t \text{ dato il contesto } w_{\{1:t-1\}}]$$

L'obiettivo di addestramento è sintattico in senso preciso: massimizzare la log-verosimiglianza dei token del corpus. Non vi è, nell'obiettivo, alcun riferimento diretto alla realtà — agli eventi che le sequenze di token descrivono, alle relazioni

causali che implicano, ai fatti del mondo che affermano. Il modello impara la struttura statistica del testo, non la struttura del mondo di cui il testo parla.

Sia  $M_{\text{realtà}}$  la struttura del dominio reale — l'insieme dei fatti, delle relazioni causali, degli eventi che un parlante competente conosce quando produce o valuta testo. La semantica del modello è l'insieme di ciò che il modello può inferire correttamente su  $M_{\text{realtà}}$  a partire dalla distribuzione appresa  $p_\theta$ . Il gap semantico è:

$$\text{Gap}(\theta, M) = H(M_{\text{realtà}}) - I(M_{\text{realtà}}; \Sigma_\theta)$$

dove  $H(M_{\text{realtà}})$  è l'entropia della struttura reale del dominio — la complessità informativa del mondo — e  $I(M_{\text{realtà}}; \Sigma_\theta)$  è l'informazione mutua tra la struttura reale e la struttura sintattica appresa dal modello. Il gap misura quanta struttura del mondo reale non è catturata dalla distribuzione statistica del testo.

**[CONGETTURA FONDATA]**

Proposizione (loss limite come misura del gap semantico — tipo C). Kaplan et al. (2020) mostrano empiricamente che la cross-entropy loss  $L(N)$  di un LLM con  $N$  parametri segue una legge di potenza:

$$L(N) \approx (N_c / N)^\alpha + L_\infty$$

dove  $N_c$  e  $\alpha$  sono costanti empiriche e  $L_\infty > 0$  è il valore limite della loss per  $N \rightarrow \infty$ . Questo valore limite non è zero — non può essere zero perché il linguaggio naturale contiene entropia irriducibile (ambiguità lessicale, riferimenti contestuali, informazione pragmatica che non è nel testo). La proposizione che  $L_\infty = \text{Gap}(\theta, M)$  nella decomposizione informativa è una congettura strutturale di tipo C: supportata dall'analisi formale ma non dimostrata nel senso pieno perché richiede ipotesi sulla struttura di  $M_{\text{realtà}}$  che non sono formalmente specificate nel lavoro empirico di Kaplan et al.

La connessione con la Famiglia IV è diretta. Il teorema di Tarski (Capitolo II) afferma che nessun sistema formale consistente può definire al proprio interno il predicato di verità per il proprio linguaggio. Il predicato di verità è la funzione semantica per eccellenza: dice, per ogni sentenza, se essa corrisponde alla realtà. Nessun LLM può imparare questa funzione dall'interno del corpus — perché il corpus è una struttura sintattica, e la funzione di verità è semantica.  $L_\infty > 0$  è la traccia empiricamente misurabile di questo risultato formale.

Le allucinazioni sono eventi in cui il modello campiona dalla regione del gap: produce testo coerente con la distribuzione sintattica appresa ma incoerente con

M\_realtà. Formalmente, sia A l'evento 'allucinazione' — il modello produce una proposizione  $\varphi$  tale che  $p_\theta(\varphi \mid \text{contesto})$  è alta ma  $\varphi$  è falsa in M\_realtà. La probabilità di allucinazione per una proposizione  $\varphi$  è:

$$P(A \mid \varphi) = P(\varphi \text{ falsa in M} \mid p_\theta(\varphi \mid c) \text{ alta})$$

Questa probabilità è positiva per costruzione, perché la distribuzione  $p_\theta$  ha supporto su sequenze che non corrispondono a nessun fatto reale. Non è una proprietà del metodo di campionamento (greedy, beam search, nucleus sampling) — è una proprietà della funzione appresa. Un modello con  $L_\infty = 0$  non allucinerebbe mai; ma un modello con  $L_\infty = 0$  non esiste per il teorema di Tarski applicato a sistemi sufficientemente espressivi.

## 2. La legge di scaling delle allucinazioni

La legge di scaling di Kaplan et al. implica una struttura specifica per la frequenza di allucinazione in funzione della dimensione del modello. Sia  $\alpha(N)$  la frequenza di allucinazione attesa per un modello di N parametri su un benchmark di fatti verificabili. Assumendo che  $\alpha(N)$  sia monotona decrescente in N e che converga allo stesso limite della loss:

$$\alpha(N) \approx \alpha_0 \cdot (N_c / N)^\beta + \alpha_\infty$$

con  $\alpha_\infty > 0$  il tasso di allucinazione limite, indipendente dalla dimensione del modello. Questo è il risultato strutturale: il tasso di allucinazione non converge a zero con l'aumento della dimensione del modello. Converge a un valore positivo  $\alpha_\infty$  determinato dal gap semantico irriducibile  $L_\infty$ .

La curva empirica di OpenAI per GPT-4 vs GPT-3.5 su benchmark di fattualità (TruthfulQA, HellaSwag, MMLU) mostra esattamente questa struttura: i modelli più grandi allucinano meno, ma la riduzione è asintotica, non lineare. I modelli della famiglia GPT-4 mostrano tassi di allucinazione nell'intorno del 15-25% su domande aperte che richiedono conoscenza fatturale precisa — un miglioramento sostanziale rispetto ai modelli precedenti, ma non una convergenza a zero. L'interpretazione strutturale di questo dato è che siamo nel regime dove il primo termine  $(N_c / N)^\beta$  è diventato piccolo ma  $\alpha_\infty$  è ancora dominante.

[CONGETTURA FONDATA]

Congettura (legge di scaling delle allucinazioni — tipo C). Per qualsiasi LLM addestrato su corpus linguistico reale, il tasso di allucinazione su proposizioni fattuali verificabili converge a un valore limite  $\alpha_\infty > 0$  per  $N \rightarrow \infty$ . Questo valore è proporzionale al gap semantico irriducibile:

$$\alpha_\infty \propto L_\infty = \text{Gap}(\Sigma_\theta, M_{\text{realtà}})$$

La congettura non è dimostrata nel senso formale pieno — richiede ipotesi sulla struttura di  $M_{\text{realtà}}$  e sulla relazione tra loss sintattica e tasso di errore semantico che sono plausibili ma non dimostrate formalmente. È tuttavia coerente con tutti i dati empirici disponibili sui modelli della famiglia GPT, LLaMA, Gemini, Claude al momento della redazione di questo testo.

### *3. Il paradigma neuro-simbolico: struttura formale della proposta*

Il rimedio proposto nel post integra il ragionamento simbolico: query su database strutturati, risolutori matematici, motori di inferenza formale. L'architettura generale di questi sistemi è la seguente. Il modulo neurale (LLM) riceve linguaggio naturale in input e produce linguaggio naturale in output, con la capacità aggiuntiva di generare chiamate a strumenti simbolici. Il modulo simbolico (database, solver) riceve query strutturate e restituisce risultati precisi. L'interfaccia tra i due livelli è bidirezionale:

$$\text{LLM (livello neurale, sintattico)} \rightleftharpoons \text{Interfaccia} \rightleftharpoons \text{DB/Solver (livello simbolico)}$$

Esempi concreti di questa architettura includono: RAG (Retrieval-Augmented Generation), in cui il LLM formula query di ricerca e riceve chunks di testo rilevanti da un database vettoriale; tool-use (function calling in GPT-4, Claude, Gemini), in cui il LLM genera chiamate a funzioni strutturate con argomenti tipizzati; sistemi code-interpreter, in cui il LLM genera codice Python che viene eseguito da un interprete deterministico; e sistemi di theorem proving assistiti da LLM (Lean copilot, Isabelle/LLM), in cui il LLM suggerisce passi di dimostrazione verificati da un proof-checker formale.

L'intuizione alla base di questa architettura è corretta nella sua struttura fondamentale: il livello simbolico fornisce un ancoraggio alla realtà che il livello neurale non può produrre dall'interno. Il database contiene fatti verificati; il solver produce risultati matematicamente corretti; il proof-checker garantisce la validità logica. Il livello semantico è esternalizzato in strutture dove il gap semantico è zero o misurabile con precisione.

Questa è precisamente la struttura della gerarchia tarskiana applicata all'architettura dei sistemi AI. Il livello neurale è  $L_0$  — il linguaggio oggetto che non può definire la propria verità. Il livello simbolico è  $L_1$  — il metalinguaggio che definisce la verità per  $L_0$ . La gerarchia di Tarski è non una metafora applicata all'architettura: è la struttura formale che l'architettura implementa.

#### *4. Perché il confine si sposta ma non scompare: tre argomenti formali*

Il paradigma neuro-simbolico è un progresso reale — riduce le allucinazioni nel dominio di copertura degli strumenti simbolici, in modo misurabile e sostanziale. Ma tre argomenti formali mostrano che non elimina il gap semantico: lo sposta, lo ridistribuisce, lo trasforma in forme diverse. Questo non è una critica al paradigma — è la sua caratterizzazione strutturalmente corretta, che permette di capire dove i limiti residui si trovano e come affrontarli.

##### **Argomento 1 — Il gap dell'interfaccia.**

L'interfaccia tra il livello neurale e il livello simbolico è essa stessa un problema di chiusura semantica. Per generare una query SQL corretta da una domanda in linguaggio naturale, l'LLM deve interpretare semanticamente la propria produzione sintattica — deve capire cosa la domanda chiede, qual è lo schema del database rilevante, quali colonne e relazioni sono pertinenti. Questa interpretazione è soggetta allo stesso gap semantico della produzione originale.

Formalmente: sia  $Q$  la funzione che mappa una query in linguaggio naturale  $q \in \text{NL}$  a una query strutturata  $q' \in \text{SQL}$ .  $Q$  deve essere semanticamente corretta —  $q'$  deve recuperare esattamente l'informazione che  $q$  chiedeva. Ma la correttezza semantica di  $Q$  è soggetta al gap:

$$\text{Gap}(Q_{\text{LLM}}, Q_{\text{ideale}}) = H(Q_{\text{ideale}}) - I(Q_{\text{ideale}}; \Sigma_{\{Q_{\text{LLM}}\}}) > 0$$

dove  $Q_{\text{ideale}}$  è la funzione di traduzione semanticamente corretta (quella che produce sempre la query strutturata che risponde alla domanda originale) e  $\Sigma_{\{Q_{\text{LLM}}\}}$  è la struttura sintattica del LLM utilizzato per la traduzione. Questo gap non è zero: l'ambiguità delle domande in linguaggio naturale, la dipendenza contestuale del riferimento, la necessità di conoscenza di sfondo non contenuta nella query — tutte queste fonti di gap persistono all'interfaccia. Le query mal formate, le proiezioni sbagliate, le condizioni WHERE errate sono le allucinazioni

dell'interfaccia: morfologicamente diverse dalle allucinazioni fattuali, strutturalmente identiche.

**[FORMALMENTE DIMOSTRATO]**

Lemma (gap dell'interfaccia come corollario di Tarski — tipo D). Sia  $S = (\text{LLM}, \text{Interface}, \text{DB})$  un sistema neuro-simbolico. Sia  $T\_S$  la funzione di traduzione implementata dall'interfaccia (dal linguaggio naturale al linguaggio delle query). Se  $T\_S$  è implementata da un LLM, allora  $T\_S$  è soggetta al teorema di Tarski: non esiste alcun componente di  $S$  che possa definire internamente il predicato di correttezza semantica di  $T\_S$  per tutte le query possibili.

Dimostrazione. Per definire la correttezza semantica di  $T\_S$  occorre un predicato  $\text{Corr}(q, q')$  che affermi 'q' risponde alla domanda q'. Ma  $\text{Corr}$  è un predicato sulla relazione tra il linguaggio delle query (sintattico) e l'intenzione dell'utente (semantica). Per il teorema di Tarski, questo predicato non può essere definito all'interno del sistema linguistico che implementa l'interfaccia — richiede un metalinguaggio.  $\square$

**Argomento 2 — Il gap della copertura.**

Il livello simbolico riduce il gap semantico soltanto nel dominio di copertura degli strumenti disponibili. Un database SQL copre la struttura dati che contiene; un solver matematico copre il dominio formale per cui è stato progettato; un motore di ricerca copre il corpus indicizzato. Per tutto ciò che è fuori dal dominio di copertura, il sistema neuro-simbolico si comporta come un LLM puro — il livello simbolico non è invocato, e il gap semantico rimane invariato.

La struttura formale è quella di una decomposizione del dominio:

$$D = D_{\text{cov}} \cup D_{\text{unc}} \quad (\text{dominio coperto e non coperto dallo strato simbolico})$$

Il tasso di allucinazione del sistema neuro-simbolico è:

$$\alpha_{\text{NS}}(N) = P(D_{\text{cov}}) \cdot \alpha_{\text{cov}}(N) + P(D_{\text{unc}}) \cdot \alpha_{\text{LLM}}(N)$$

dove  $\alpha_{\text{cov}}(N)$  è il tasso di allucinazione nel dominio coperto (molto ridotto: dipende dalla correttezza dell'interfaccia, non dal gap semantico del LLM),  $\alpha_{\text{LLM}}(N)$  è il tasso di allucinazione del LLM puro (che converge a  $\alpha_{\infty} > 0$ ), e  $P(D_{\text{cov}})$ ,  $P(D_{\text{unc}})$  sono le probabilità che una query cada nel dominio coperto o non coperto. Per query nel dominio coperto il sistema funziona come promesso. Per query fuori dal dominio coperto il gap semantico è invariato.

Questo ha una conseguenza non ovvia: il sistema neuro-simbolico ha prestazioni superiori all'LLM puro nel dominio coperto, ma può avere prestazioni inferiori per

query che cadono al bordo tra  $D_{cov}$  e  $D_{unc}$  — dove il sistema deve decidere se invocare lo strumento simbolico o rispondere direttamente. Questa decisione è soggetta al gap dell'interfaccia: il LLM può invocare lo strumento sbagliato, invocare lo strumento corretto con la query sbagliata, o non invocare alcuno strumento quando dovrebbe. Il confine tra  $D_{cov}$  e  $D_{unc}$  è una nuova sorgente di gap semantico creata dall'architettura neuro-simbolica stessa.

### Argomento 3 — La regressione del grounding.

L'argomento più profondo è strutturale e non dipende da considerazioni ingegneristiche specifiche. Si chiama regressione del grounding: ogni livello simbolico che fornisce semantica al livello neurale è esso stesso un sistema formale che richiede interpretazione semantica. Un database SQL contiene fatti strutturati — ma la struttura dello schema (quali tabelle, quali colonne, quali relazioni) è arbitraria e richiede interpretazione. Un motore di ricerca restituisce documenti rilevanti — ma la rilevanza è una proprietà semantica che dipende dall'intenzione della ricerca. Un proof-checker verifica la correttezza formale — ma la corrispondenza tra la proposizione formale e l'affermazione in linguaggio naturale che si vuole dimostrare è soggetta al gap di traduzione.

Formalmente: sia  $S_n$  il sistema al livello  $n$  della gerarchia neuro-simbolica ( $S_0 = \text{LLM}$ ,  $S_1 = \text{strumento simbolico}$ ,  $S_2 = \text{verifica dello strumento simbolico}$ , ...). Il gap semantico al livello  $n$  è:

$$\text{Gap}_n = H(M_n) - I(M_n; \Sigma_n)$$

dove  $M_n$  è il dominio semantico del livello  $n$  e  $\Sigma_n$  è la struttura sintattica del sistema al livello  $n$ . La regressione del grounding afferma che  $\text{Gap}_n > 0$  per ogni  $n$  finito. La dimostrazione segue direttamente dal teorema di Tarski: per ogni sistema  $S_n$  sufficientemente espressivo da permettere auto-riferimento (e qualsiasi sistema pratico lo è), il predicato di verità semantica non è definibile all'interno di  $S_n$ .

#### [FORMALMENTE DIMOSTRATO]

Teorema della regressione del grounding (tipo D, corollario di Tarski). Per qualsiasi sistema neuro-simbolico stratificato  $S = (S_0, S_1, \dots, S_k)$  di profondità finita  $k$ , esiste un gap semantico residuo:

$$\text{Gap}_{\text{residuo}}(S) = H(M_{\text{realtà}}) - I(M_{\text{realtà}}; \bigcup_{i=0}^k \Sigma_i) > 0$$

cioè la struttura sintattica composta di tutti i livelli del sistema non cattura completamente la struttura semantica del dominio reale.

Dimostrazione. Per induzione su  $k$ . Base ( $k=0$ ): il gap è  $\text{Gap}(\Sigma_{\text{LLM}}, M_{\text{realtà}}) > 0$  per il teorema di Tarski applicato al livello neurale. Passo induttivo: assumiamo  $\text{Gap}_k > 0$ . Il sistema  $S_{\{k+1\}}$  aggiunge struttura sintattica  $\Sigma_{\{k+1\}}$  a quella esistente. L'informazione mutua  $I(M_{\text{realtà}} ; \bigcup_{i \leq k+1} \Sigma_i) \leq H(M_{\text{realtà}})$  (per la subadditività dell'entropia). L'uguaglianza richiederebbe che la struttura sintattica composta catturasse completamente  $M_{\text{realtà}}$  — impossibile per il teorema di Tarski per sistemi con auto-riferimento.  $\square$

### 5. La misura quantitativa del miglioramento: navigazione della frontiera

I tre argomenti mostrano che il paradigma neuro-simbolico non elimina il gap semantico. Ma mostrano anche perché è un progresso reale: riduce il gap nel dominio coperto in modo strutturalmente controllabile. Il miglioramento è misurabile in termini precisi.

Sia  $\Delta\alpha(\text{NS}, \text{LLM}) = \alpha_{\text{LLM}} - \alpha_{\text{NS}}$  il miglioramento nel tasso di allucinazione del sistema neuro-simbolico rispetto all'LLM puro. Dalla decomposizione precedente:

$$\Delta\alpha = P(D_{\text{cov}}) \cdot (\alpha_{\text{LLM}(N)} - \alpha_{\text{cov}(N)})$$

Il miglioramento è proporzionale a due fattori: la copertura del dominio simbolico  $P(D_{\text{cov}})$  e la riduzione del tasso di allucinazione all'interno del dominio coperto  $\alpha_{\text{LLM}} - \alpha_{\text{cov}}$ . Il secondo fattore è tipicamente grande: nel dominio coperto,  $\alpha_{\text{cov}}$  può essere vicino a zero (una query SQL eseguita correttamente restituisce la risposta giusta con probabilità 1 meno la probabilità di errore dell'interfaccia). Il primo fattore — la copertura — determina quanto del miglioramento si traduce in riduzione globale.

Questa analisi ha implicazioni pratiche immediate per la progettazione di sistemi AI. Il parametro critico di un sistema neuro-simbolico non è la potenza del modello neurale né la precisione del modulo simbolico: è la copertura  $P(D_{\text{cov}})$  e la qualità dell'interfaccia  $\text{Gap}(Q_{\text{LLM}}, Q_{\text{ideale}})$ . I sistemi che massimizzano  $P(D_{\text{cov}})$  attraverso una progettazione deliberata del dominio coperto ottengono il massimo vantaggio dall'architettura neuro-simbolica. I sistemi che hanno interfacce di alta qualità (basso gap di traduzione) minimizzano il tasso di allucinazione nel dominio coperto.

La legge di Amdahl dell'allucinazione. Il miglioramento massimo ottenibile da un sistema neuro-simbolico è limitato dalla percentuale del dominio non coperta dallo strato simbolico. Se  $P(D_{unc}) = p$ , il tasso di allucinazione del sistema neuro-simbolico è almeno  $p \cdot \alpha_{\infty}$  — il tasso di allucinazione irriducibile moltiplicato per la frazione di dominio non coperta. Per ridurre il tasso di allucinazione al di sotto di  $p \cdot \alpha_{\infty}$  è necessario espandere la copertura simbolica, non aumentare la dimensione del modello neurale.

$$\alpha_{NS} \geq P(D_{unc}) \cdot \alpha_{\infty} \quad [\text{legge di Amdahl dell'allucinazione}]$$

**[CONGETTURA FONDATA]**

Congettura (legge di Amdahl dell'allucinazione — tipo C). Il limite inferiore al tasso di allucinazione di un sistema neuro-simbolico è:

$$\alpha_{NS} \geq P(D_{unc}) \cdot \alpha_{\infty}$$

dove  $P(D_{unc})$  è la probabilità che una query cada fuori dal dominio coperto dallo strato simbolico e  $\alpha_{\infty}$  è il tasso di allucinazione limite del modello neurale. Per ridurre  $\alpha_{NS}$  a  $\varepsilon$  occorre o espandere  $D_{cov}$  ( $P(D_{unc}) \leq \varepsilon/\alpha_{\infty}$ ) o ridurre  $\alpha_{\infty}$  (impossibile per dimensionamento del modello, come mostrato dalla legge di scaling). La congettura è di tipo C perché dipende dall'ipotesi che  $\alpha_{cov} \approx 0$  nel dominio coperto, che è approssimativamente vera per sistemi simbolici deterministici ma non è garantita per interfacce di traduzione imperfette.

## 6. Il ragionamento neuro-simbolico come paradossologia costruttiva

La struttura dell'argomento fino a questo punto potrebbe sembrare pessimistica: il gap semantico è irriducibile, la regressione del grounding è inevitabile, le allucinazioni persistono. Ma questa lettura ignora esattamente ciò che il Capitolo XI ha sviluppato: le impossibilità strutturali non sono ostacoli del pensiero ma condizioni della sua possibilità. Il paradigma neuro-simbolico è il caso più chiaro di paradossologia costruttiva nell'AI contemporanea.

Il pattern è preciso. Il gap semantico della Famiglia IV — l'impossibilità di chiusura semantica del solo livello neurale — ha la forma specifica di una distanza tra distribuzione di token e struttura del mondo. Questa forma specifica ha determinato la struttura precisa della soluzione: il paradigma neuro-simbolico separa i due livelli (neurale e simbolico) lungo esattamente il confine identificato dal gap. L'architettura non è casuale: ha la geometria del confine che incontra. Il livello neurale gestisce ciò che è naturalmente sintattico (comprensione del

linguaggio naturale, generazione di testo, ragionamento analogico). Il livello simbolico gestisce ciò che è naturalmente semantico (verifica dei fatti, calcolo, inferenza formale). La divisione del lavoro tra i due livelli replica la struttura della distinzione tarskiana.

Questo non è un accidente ingegneristico: è la forma costruttiva che il confine sintassi-semantica produce quando viene incontrato da un sistema sufficientemente ricco e con sufficiente persistenza operativa (le quattro condizioni del Principio della Costruttività Condizionata del Capitolo XI). La struttura della gerarchia tarskiana —  $L_0$  non può definire la propria verità ma  $L_1$  può definire la verità di  $L_0$  — è esattamente la struttura implementata dall'architettura neuro-simbolica: il LLM ( $L_0$ ) non può verificare le proprie affermazioni, ma il database o il proof-checker ( $L_1$ ) può verificarle.

Il surplus generativo è reale e misurabile. Rispetto all'LLM puro, il sistema neuro-simbolico produce una struttura genuinamente nuova — una struttura la cui forma è determinata dalla forma del confine stesso. E questa struttura genera a sua volta nuovi confini (il gap dell'interfaccia, il gap della copertura) che producono nuove pressioni costruttive verso soluzioni ulteriori: l'ottimizzazione automatica dello schema del database, l'apprendimento della funzione di traduzione, la verifica formale dell'interfaccia. Ogni soluzione genera il proprio confine; ogni confine genera la propria soluzione parziale. La gerarchia tarskiana non è un ostacolo: è il motore della progressione.

#### [PROPOSTA INTERPRETATIVA]

Proposta interpretativa (il paradigma neuro-simbolico come istanza della paradossologia costruttiva — tipo I). Il pattern storico di sviluppo dei sistemi AI — da LLM puri a sistemi con RAG, a sistemi con tool-use, a sistemi con proof-checking, verso sistemi con architetture neurali-simboliche sempre più profonde — segue il pattern dell'ascesa gödeliana (Capitolo XI): ogni livello della gerarchia supera il limite del livello precedente generando struttura determinata dalla forma del limite. Il surplus generativo  $\Gamma(\text{Impossibilità Famiglia IV, Sistema AI}) = K(\text{Architettura NS} \mid \text{LLM}) > 0$  è misurabile in termini di complessità della struttura aggiunta.

Questa è una proposta interpretativa di tipo I: cattura strutture reali del processo di sviluppo dell'AI, ma non è una derivazione formale dell'architettura neuro-simbolica dalla logica dell'impossibilità di chiusura semantica.

### *7. Implicazioni per la ricerca: dove investire, dove non aspettarsi progressi*

La formalizzazione del gap semantico e della struttura dei sistemi neuro-simbolici ha implicazioni concrete per la direzione della ricerca. Tre indicazioni emergono direttamente dall'analisi.

Prima indicazione: l'interfaccia è il collo di bottiglia principale. La legge di Amdahl dell'allucinazione mostra che il tasso di allucinazione del sistema è dominato dalla qualità dell'interfaccia (Gap dell'interfaccia) e dalla copertura del dominio simbolico. Il miglioramento del modello neurale — dimensione, architettura, dati di addestramento — ha rendimenti marginali decrescenti rispetto al gap semantico irriducibile. Il miglioramento dell'interfaccia ha rendimenti non decrescenti: un'interfaccia perfetta ( $\alpha_{\text{cov}} = 0$ ) nel dominio coperto trasforma il sistema in un sistema senza allucinazioni per quel dominio. La ricerca dovrebbe concentrarsi sulla traduzione neurale-simbolica, sulla verifica automatica dell'interfaccia, e sull'espansione controllata della copertura.

Seconda indicazione: la copertura simbolica deve essere progettata, non emergente. Il dominio coperto  $D_{\text{cov}}$  di un sistema neuro-simbolico non è definito dall'LLM — è definito dagli strumenti simbolici disponibili e dal modo in cui l'LLM apprende a invocarli. La progettazione deliberata di  $D_{\text{cov}}$  — identificare i sotto-domini dove le allucinazioni hanno il costo maggiore (sicurezza, medicina, diritto, matematica) e costruire strumenti simbolici con alta copertura per quei sotto-domini — è la strategia più efficiente rispetto alla formalizzazione qui sviluppata. I sistemi AI per applicazioni critiche dovrebbero avere  $\alpha_{\text{cov}} \approx 0$  per il loro sotto-dominio di applicazione e  $P(D_{\text{unc}}) \approx 0$  per lo stesso sotto-dominio — non  $\alpha_{\text{LLM}}$  basso su benchmark generali.

Terza indicazione: non attendersi l'eliminazione del gap su domini aperti. Per sistemi AI destinati a domini aperti — rispondere a qualsiasi domanda in linguaggio naturale su qualsiasi argomento — il gap semantico è irriducibile e la legge di Amdahl implica che le allucinazioni persistono. La strategia corretta non è eliminare le allucinazioni ma calibrare l'incertezza: produrre sistemi che sappiano quando si trovano nel dominio coperto (dove possono essere precisi) e quando si trovano nel dominio non coperto (dove devono comunicare incertezza). La calibrazione dell'incertezza — la corrispondenza tra la confidenza espressa dal modello e la probabilità di correttezza effettiva — è un problema di Famiglia IV:

richiederebbe che il modello definisse internamente il proprio predicato di affidabilità, soggetto alla stessa regressione del grounding. Ma la calibrazione approssimata è raggiungibile, ed è la versione costruttiva della risposta all'impossibilità.

---

# Capitolo XI

## La paradossologia costruttiva

Come le impossibilità strutturali generano il pensiero che le incontra, e perché il confine è la condizione della novità

### Mappa epistemica — Capitolo XI

| Status          | Cod. | Claim                                                                                                                                                                           |
|-----------------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | L'ascesa gödeliana: la sequenza $T_0 \times T_1 \times T_2 \times \dots$ è strettamente crescente con ordinali di prova crescenti (Gentzen 1936).                               |
| DIMOSTRATO      | D    | Gli ordinali di prova e la loro sequenza strettamente crescente sono formalmente dimostrati.                                                                                    |
| CONGETTURA      | C    | Il surplus generativo $\Gamma(I, S) = K(S' S)$ come misura della costruttività: definizione precisa ma K non è calcolabile.                                                     |
| INTERPRETAZIONE | I    | La creatività come navigazione dei punti fissi dell'impossibile; la paradossologia costruttiva nei sistemi biologici: proposte interpretative con forte supporto argomentativo. |

### I. Il rovesciamento necessario

Il ciclo ha costruito, nelle dieci sezioni precedenti, una teoria delle impossibilità strutturali. Ha identificato cinque famiglie, le ha derivate da un principio unificante, ha mostrato come esse si manifestino nei sistemi formali, nei sistemi biologici, nei sistemi cognitivi e nelle istituzioni della scienza. La tassonomia è precisa, le famiglie sono genuine, i teoremi che le fondano sono dimostrati nel senso tecnico pieno o argomentati con rigore sufficiente a distinguerli dalle mere congetture. Se ci si fermasse qui, il ciclo sarebbe una mappa dei confini — un atlante di ciò che nessun sistema può fare.

Ma questo sarebbe anche, nella sua struttura profonda, un errore interpretativo. L'errore non riguarda i teoremi — quelli restano — ma la lettura che di essi si trae. La tentazione di leggere una teoria delle impossibilità strutturali come una teoria dei limiti del pensiero è reale e comprensibile, ma è sbagliata, e lo è in un modo che ha conseguenze pratiche per chiunque voglia usare la teoria come strumento invece che come commemorazione.

La tesi di questo capitolo è che le impossibilità strutturali non sono gli ostacoli del pensiero, ma la sua condizione di possibilità. Questa affermazione non è una consolazione filosofica, né una reinterpretazione ottimistica di fatti spiacevoli. È un enunciato con struttura formale precisamente formulabile, dimostrabile nei casi in cui si dispone di strumenti adeguati e argomentabile con solidità nei casi in cui gli strumenti formali mancano ancora. L'enunciato afferma qualcosa di specifico: ogni sistema che incontra il confine tra la propria struttura sintattica e la struttura semantica del proprio dominio — ogni sistema che si scontra con la forma precisa della propria impossibilità — produce strutture che non erano prevedibili dalla sua struttura precedente, strutture la cui forma è determinata dalla forma del confine stesso. Il confine non è neutro rispetto al pensiero che lo incontra: è informativo. Contiene, nella geometria della propria curvatura, l'indicazione della direzione in cui il sistema deve espandersi per abitare il territorio che il confine delimita. Questa non è metafora: si formalizza con precisione attraverso il concetto di surplus generativo, si misura con la complessità condizionata di Kolmogorov, e si dimostra per la famiglia più importante attraverso l'ascesa gödeliana come sequenza ordinale di teorie di forza dimostrativa strettamente crescente.

Il nome che propongo per questo fenomeno — la paradossologia costruttiva — è deliberatamente provocatorio. I paradossi fondativi del Novecento hanno il nome sbagliato. Il paradosso di Russell non è un paradosso nel senso di una contraddizione da cui non si esce: è il germe della teoria dei tipi, della teoria degli insiemi assiomatica, di tutti i fondamenti moderni della matematica. La sentenza gödeliana non è un vicolo cieco del pensiero formale: è il motore che ha generato la gerarchia delle teorie aritmetiche, la teoria degli ordinali di prova, l'analisi ordinale come campo. Il problema dell'arresto di Turing non è una barriera alla computazione: è il punto di partenza dell'intera teoria della computabilità e dei gradi di indecidibilità. In ciascuno di questi casi — e in molti altri nei domini biologici e cognitivi che verranno analizzati nelle sezioni seguenti — la struttura precisa dell'impossibilità ha determinato la struttura precisa della teoria che l'impossibilità ha reso necessaria. Non c'è nulla di accidentale in questa determinazione: è il meccanismo della paradossologia costruttiva, e comprenderlo è l'obiettivo di questo capitolo.

Prima, però, è necessaria una distinzione che il carattere entusiastico dell'affermazione centrale potrebbe oscurare. Non tutte le impossibilità strutturali

sono costruttive. Alcune lo sono, alcune non lo sono, e la differenza tra le due categorie non dipende dall'impossibilità in sé, ma dal sistema che la incontra e dalle condizioni in cui l'incontro avviene. Una teoria delle impossibilità costruttive che non distinguesse con cura tra i casi in cui la costruttività si attiva e quelli in cui fallisce sarebbe intellettualmente disonesta — esattamente il tipo di ottimismo mal fondato che una teoria dei limiti dovrebbe evitare. Questa distinzione — il principio della costruttività condizionata — occupa la sezione centrale dell'analisi, e non viene affrontata dopo aver già convinto il lettore della tesi generale, ma prima: è parte dell'argomento, non un'aggiunta difensiva.

## II. L'ascesa gödeliana: meccanismo e struttura

Il meccanismo fondamentale della paradossologia costruttiva, quello per cui la dimostrazione è più rigorosa, è l'ascesa gödeliana. Vale la pena svilupparlo con la precisione tecnica che merita, perché la sua struttura rivelerà il pattern che si ripresenta, in forme meno formali ma riconoscibili, negli altri domini.

Sia  $T_0$  una teoria del primo ordine ricorsivamente assiomatizzabile, consistente, e sufficientemente ricca da contenere una codifica dell'aritmetica elementare — il che significa, in concreto, che  $T_0$  contiene almeno PA (l'aritmetica di Peano) o una sua estensione. Per il primo teorema di incompletezza di Gödel (1931), esiste una sentenza  $G_{\{T_0\}}$  tale che  $T_0 \not\models G_{\{T_0\}}$  e  $T_0 \not\models \neg G_{\{T_0\}}$ : la sentenza è indecidibile in  $T_0$ . La costruzione di  $G_{\{T_0\}}$  usa il lemma della diagonalizzazione, che produce esplicitamente una sentenza che dice di se stessa di essere indimostrabile nella teoria. Nella semantica standard dei numeri naturali — la struttura  $\mathbb{N}$  con le operazioni usuali —  $G_{\{T_0\}}$  è vera: essa afferma la propria indimostrabilità in  $T_0$ , e questa affermazione è vera perché  $T_0$  è consistente. L'ind decidibilità non è un'oscurità semantica: è un fatto matematico preciso sulla distanza tra la struttura dimostrativa di  $T_0$  e la struttura semantica standard.

Fin qui il teorema di incompletezza nella lettura usualment ricevuta, quella che vede in  $G_{\{T_0\}}$  un limite. Ma si consideri ciò che accade quando si risponde all'impossibilità invece di limitarsi a constatarla. Si definisce  $T_1 = T_0 \cup \{G_{\{T_0\}}\}$  — la teoria ottenuta aggiungendo  $G_{\{T_0\}}$  come assioma. Tre osservazioni su  $T_1$ . Prima:  $T_1$  è consistente, perché  $G_{\{T_0\}}$  è vera nella semantica standard e qualsiasi teoria i cui assiomi sono veri nella semantica standard è consistente (se  $T_1$  fosse

inconsistente, permetterebbe di derivare  $\neg G_{-}\{T_0\}$ , ma  $\neg G_{-}\{T_0\}$  è falsa nella semantica standard — contraddizione). Seconda:  $T_1 \vdash G_{-}\{T_0\}$  per costruzione, mentre  $T_0 \not\vdash G_{-}\{T_0\}$  —  $T_1$  è strettamente più forte di  $T_0$  nel senso che dimostra qualcosa che  $T_0$  non può dimostrare. Terza:  $T_1$  è anch'essa soggetta al teorema di incompletezza, perché è anch'essa ricorsivamente assiomatizzabile, consistente, e contiene l'aritmetica. Esiste quindi  $G_{-}\{T_1\}$  — la sentenza gödeliana di  $T_1$  — che è indecidibile in  $T_1$ . Il processo si itera:  $T_2 = T_1 \cup \{G_{-}\{T_1\}\}$ , e così via.

La sequenza  $T_0 \subsetneq T_1 \subsetneq T_2 \subsetneq \dots$  è strettamente crescente nel senso delle teorie — ogni teoria dimostra strettamente più della precedente. Estendendo transfinitamente lungo gli ordinali costruttivi, si ottiene  $T_\omega = \bigcup_{\alpha < \omega} T_\alpha$ , poi  $T_{\omega+1} = T_\omega \cup \{G_{-}\{T_\omega\}\}$ , e così avanti. La gerarchia delle teorie è indicizzata dagli ordinali costruttivi, e ciascun livello è generato dall'impossibilità diagonale del livello precedente.

Questa gerarchia non è un'astrazione priva di contenuto concreto. È misurabile attraverso il concetto di ordinale di prova, denotato  $|T|$ : il più piccolo ordinale che non è provabilmente ben-ordinato in  $T$  nel senso della derivazione transfinita (la precisazione si deve a Gentzen 1936, Schütte, Feferman). I valori per le teorie principali sono:

$|PA| = \varepsilon_0$ , il primo punto fisso della funzione  $\alpha \mapsto \omega^\alpha$ , cioè il limite della torre  $\omega, \omega^\omega, \omega^{\omega^\omega}, \dots$ . Il significato è che PA può dimostrare la buona fondazione di ogni ordinale  $\alpha < \varepsilon_0$  ma non di  $\varepsilon_0$  stesso.

$|ATR_0| = \Gamma_0$ , il numero di Feferman-Schütte, primo punto fisso della funzione  $\alpha \mapsto \varphi(\alpha, 0)$  dove  $\varphi$  è la funzione di Veblen.  $ATR_0$  è la sottoteoria dell'analisi del secondo ordine che aggiunge alle comprensioni aritmetiche la ricorsione transfinita sugli insiemi di numeri naturali.

$|\Pi^1_1\text{-}CA_0| = \psi(\Omega^\omega)$  nell'analisi di collasso ordinale di Buchholz, un ordinale immensamente più grande di  $\Gamma_0$ , raggiunto attraverso la comprensione  $\Pi^1_1$  — la capacità di formare insiemi definiti da formule con un quantificatore universale del secondo ordine.

Il punto che la tabella degli ordinali rende visibile è che il pascapitolo da PA ad  $ATR_0$  non è un'operazione arbitraria: è il risultato dell'ascesa gödeliana iterata fino al livello  $\omega$  nella numerazione naturale degli assiomi, e il pascapitolo da  $ATR_0$  a

$\Pi_1\text{-CA}_0$  riflette un cambio qualitativo nel tipo di auto-riferimento che il sistema può esprimere — il salto dalla ricorsione transfinita sui predicati aritmetici alla comprensione di classi definite da proprietà del secondo ordine. In ciascuno di questi passaggi, la forma specifica dell'impossibilità al livello precedente determina la forma specifica della struttura superiore: non si ottiene una struttura qualunque, ma esattamente quella struttura la cui potenza espressiva è necessaria e sufficiente per vedere  $G_{\{T_{\text{precedente}}\}}$  come oggetto del dominio.

Questo è il nucleo della tesi: le impossibilità strutturali non dicono soltanto "fin qui e non oltre" — dicono "esattamente qui la struttura del problema richiede esattamente questo tipo di estensione per poter essere trattato." La sentenza  $G_{\{PA\}}$  non è dimostrabile in PA ma è dimostrabile in  $PA + \text{Con}(PA)$ , e questa estensione non è arbitraria: è la minima estensione che rende  $G_{\{PA\}}$  accessibile, e la natura di questa minima estensione — l'aggiunta della consistenza di PA come assioma — riflette il fatto che l'impossibilità di dimostrare  $G_{\{PA\}}$  in PA dipende esattamente dall'impossibilità di PA di vedere la propria consistenza dall'interno.

Prima di procedere agli altri domini, è opportuno formalizzare la struttura del meccanismo in modo da poterla riconoscere quando riappare in forme diverse. Il surplus generativo di un'impossibilità I per un sistema S è:

$$\Gamma(I, S) = K(S' \mid S)$$

dove  $S'$  è il sistema generato dall'incontro con I — il metalinguaggio, la teoria estesa, la nuova struttura — e  $K(S' \mid S)$  è la complessità condizionata di Kolmogorov: la lunghezza del programma più corto che produce  $S'$  avendo S come oracolo. Questa misura è zero se e soltanto se  $S'$  è completamente ricostruibile da S attraverso un algoritmo finito — se  $S'$  non porta informazione genuinamente nuova. Per l'ascesa gödeliana,  $K(T_1 \mid T_0)$  è almeno pari alla complessità di Kolmogorov di  $G_{\{T_0\}}$ , che è positiva perché  $G_{\{T_0\}}$  è un oggetto matematico definito ma non banale — non è generabile da  $T_0$  da sola attraverso nessuna procedura algoritmica finita (se lo fosse, sarebbe dimostrabile in  $T_0$  per derivazione). Il surplus generativo dell'incompletezza gödeliana è quindi positivo e misurabile.

Si osservi che la definizione di  $\Gamma$  usando la complessità condizionata di Kolmogorov ha un costo: K non è calcolabile nel senso della macchina di Turing (il teorema di Rice implica che non esiste un algoritmo che calcoli K per input

arbitrario).  $\Gamma(I, S)$  è quindi una misura teorica, non computazionale — un oggetto matematico ben definito ma non algoritmicamente accessibile. Questo non ne riduce il valore teorico: serve a misurare la costruttività in modo preciso, non a calcolarla in pratica.

### III. Il paradosso di Russell e la costruttività del collasso fondazionale

Il paradosso di Russell — scoperto nel 1902 e comunicato a Frege in quella lettera che è tra le più devastanti della storia della matematica — è il caso in cui la costruttività di un'impossibilità si manifesta con la massima spettacolarità storica, anche se con minore rigore formale rispetto all'ascesa gödeliana. Vale la pena svilupparlo in dettaglio perché rivela dimensioni della paradossologia costruttiva che l'argomento puramente logico non mostra.

La situazione ante-paradosso è la seguente. Frege aveva costruito, nei *Grundgesetze der Arithmetik* (1893, 1903), un sistema di logica pura destinato a derivare tutta l'aritmetica da principi logici soli. Il sistema includeva la Legge Base V: per ogni proprietà  $F$ , esiste l'estensione  $\{x : F(x)\}$  — l'insieme degli oggetti che soddisfano  $F$ . La legge appariva innocente, quasi ovvia: se si può caratterizzare una proprietà, si può formare l'insieme degli oggetti che la soddisfano. Non c'è nulla nell'intuizione geometrica o aritmetica che la suggerisca problematica.

Russell considera la proprietà  $F(x) = "x \text{ non appartiene a } x"$  e forma l'insieme  $R = \{x : x \notin x\}$ . La domanda se  $R \in R$  genera la contraddizione: se  $R \in R$ , allora  $R$  soddisfa la condizione definitoria di  $R$ , cioè  $R \notin R$ ; se  $R \notin R$ , allora  $R$  soddisfa la condizione, cioè  $R \in R$ . Il sistema di Frege è inconsistente.

La lettura di Frege è quella del crollo. La sua risposta a Russell, scritta mentre il secondo volume dei *Grundgesetze* era già in stampa, è uno dei momenti più desolati della letteratura matematica: "Hardly anything more unwelcome can befall a scientific writer than to have one of the foundations of his edifice shaken after the work is finished." La fondazione era effettivamente crollata. La Legge Base V non era salvabile nella forma in cui era stata formulata.

Ma ciò che è avvenuto nei vent'anni successivi non è il nulla che segue un crollo: è una delle più straordinarie fioritura intellettuali della storia della matematica.

Whitehead e Russell, lavorando a Principia Mathematica (1910-1913), costruiscono la teoria ramificata dei tipi: gli oggetti sono stratificati in una gerarchia di tipi (individui, proprietà di individui, proprietà di proprietà di individui, ...), e la formazione di insiemi è permessa soltanto all'interno di ogni tipo. La proprietà "x non appartiene a x" diventa mal tipata — R non può appartenere a se stesso perché appartiene a un tipo superiore a quello degli oggetti su cui è definita la proprietà di appartenenza che lo costituisce. Il paradosso scompare non perché sia risolto ma perché la struttura che lo generava — la possibilità di predicare l'appartenenza a se stessi — è resa illegale dalla grammatica del sistema.

Zermelo, lavorando in parallelo, prende una strada diversa ma convergente: invece di stratificare gli oggetti in tipi, restringe la formazione degli insiemi al principio di separazione (un insieme si forma come sottoinsieme di un insieme già dato, non come raccolta arbitraria di tutti gli oggetti che soddisfano una proprietà). Il sistema ZF (Zermelo 1908, poi esteso da Fraenkel) è il fondamento dell'analisi matematica moderna. von Neumann porta un'ulteriore variazione: la distinzione tra insiemi e classi proprie (come V, la classe di tutti gli insiemi, che è troppo grande per essere essa stessa un insieme). Queste tre risposte — la teoria dei tipi, ZF, la teoria degli insiemi con classi — sono strutturalmente diverse, incompatibili in certi dettagli, e tutte e tre genuine produzioni intellettuali che non esistevano prima del paradosso.

Il surplus generativo  $K(\text{ZF} + \text{teoria\_dei\_tipi} + \text{NBG} \mid \text{Logica\_naïve\_di\_Frege})$  è enorme — queste sono teorie matematiche di complessità descrittiva molto superiore alla Legge Base V che ha generato il collasso. La domanda teoricamente precisa è: perché il collasso ha prodotto queste specifiche strutture e non altre? Perché la risposta al paradosso di Russell non è stata "la matematica è impossibile" o "gli insiemi non esistono"?

La risposta è nella struttura dell'impossibilità stessa. Il paradosso di Russell non è un'impossibilità generica — è un'impossibilità con forma precisa: emerge dalla auto-appartenenza, dalla possibilità di applicare il predicato di non-appartenenza a se stessi. La struttura delle soluzioni riflette esattamente la forma di questa impossibilità: tutte le soluzioni impongono una limitazione sulla auto-applicazione dei predicati, ciascuna attraverso un meccanismo diverso (la stratificazione dei tipi, la separazione all'interno di insiemi dati, la distinzione tra insiemi e classi) ma con

lo stesso obiettivo formale. La geometria del confine ha determinato la geometria delle strutture oltre il confine.

Un dettaglio storico che la ricostruzione standard spesso omette: Poincaré, in contemporanea, sviluppa la nozione di impredicatività — la proprietà di una definizione che fa riferimento alla totalità di cui l'oggetto definito fa parte. Il paradosso di Russell è una definizione impredicativa ( $R$  è definito come l'insieme di tutti gli insiemi con certa proprietà, e poi ci si chiede se  $R$  stesso ha quella proprietà). L'interdizione delle definizioni impredicative è un'altra risposta al paradosso, incompatibile con parti di ZF ma necessaria per certi programmi predicativi in filosofia della matematica. Anche questa è una struttura generata dall'impossibilità — ancora con forma determinata dalla forma della impossibilità stessa.

#### IV. Cantor e la struttura generativa degli insiemi transfiniti

Il caso di Cantor precede cronologicamente il paradosso di Russell e ha struttura più chiara dal punto di vista formale, anche se la sua costruttività è meno ovvia a prima vista perché l'impossibilità cantoriana non ha la forma di una contraddizione ma di una distanza metrica.

Nel 1891 Cantor dimostra che per qualsiasi insieme  $A$ , l'insieme delle parti  $P(A)$  ha cardinalità strettamente maggiore di quella di  $A$ : non esiste una biiezione tra  $A$  e  $P(A)$ . La dimostrazione usa l'argomento diagonale già descritto nel testo. Per  $A = \mathbb{N}$ , il risultato afferma che  $|P(\mathbb{N})| > |\mathbb{N}|$ , cioè che i sottoinsiemi dei numeri naturali non sono numerabili. Identificando  $P(\mathbb{N})$  con i reali (ogni sottoinsieme corrisponde alla sua funzione caratteristica, che è una sequenza binaria, che è un numero reale in  $[0,1]$  in rappresentazione binaria), si ottiene  $|\mathbb{R}| > |\mathbb{N}|$ .

Ma il risultato non afferma soltanto che  $\mathbb{R}$  è "più grande" di  $\mathbb{N}$  — nel linguaggio ante-cantoriano non esistevano gli strumenti per dire cosa significasse. L'impossibilità ha forzato la costruzione di quelli strumenti. Prima del 1874 (data della prima dimostrazione di Cantor della non-numerabilità dei reali), il concetto di "grandezza" di un insieme infinito era intuitivo e mal definito. Parlare di insiemi infiniti equivaleva a evocare l'infinito come quantità unica e indivisibile — una metafisica, non una matematica. L'impossibilità di stabilire una biiezione tra  $\mathbb{N}$  e  $\mathbb{R}$  ha reso necessario un concetto preciso di cardinalità: due insiemi hanno la stessa

cardinalità se e soltanto se esiste una biiezione tra di essi. Senza la dimostrazione dell'impossibilità di tale biiezione per certi piani, la nozione stessa di cardinalità non avrebbe avuto ragione di essere costruita.

Una volta costruita la nozione di cardinalità, il teorema di Cantor  $|P(A)| > |A|$  afferma che esiste una sequenza strettamente crescente di cardinalità infinite:  $|\mathbb{N}| < |P(\mathbb{N})| < |P(P(\mathbb{N}))| < \dots$ . La gerarchia delle cardinalità transfinite —  $\aleph_0, \aleph_1, \aleph_2, \dots$  — è la struttura prodotta dall'iterazione dell'impossibilità. E con la gerarchia delle cardinalità viene la gerarchia degli ordinali transfiniti, che è più ricca:  $\omega, \omega+1, \omega \cdot 2, \omega^2, \omega\omega, \varepsilon_0, \dots$ . La teoria degli ordinali di Cantor — uno degli edifici matematici più elaborati del Novecento, con le sue regole di addizione non commutativa, moltiplicazione, potenza, il lemma di Zorn, il teorema ben-ordinamento — non era concepibile senza il confine che l'argomento diagonale ha tracciato. Il surplus generativo  $K(\text{Teoria\_ordinali} \mid \text{Matematica\_pre-cantoriana})$  è misurabilmente enorme.

Il meccanismo ha un'ulteriore struttura che vale la pena articolare. L'argomento diagonale di Cantor costruisce esplicitamente l'oggetto che testimonia l'impossibilità: la sequenza  $d$  che differisce dall' $n$ -esimo elemento nell' $n$ -esima posizione. Questo oggetto — la diagonale — non è soltanto la prova dell'impossibilità: è il primo rappresentante esplicito della cardinalità superiore. La diagonale è il germe della struttura nuova. Nella terminologia del teorema di Lawvere (Appendice A),  $d$  è la funzione diagonale  $d(a) = f(\varphi_a(a))$  che non appartiene all'immagine di  $\varphi$  — e la sua non-appartenenza è il fatto che genera la distinzione tra cardinalità. Questa struttura — l'oggetto diagonale come primo elemento della struttura superiore — si ripresenta in tutti i teoremi della Famiglia I, ed è il pattern caratterizzante della costruttività in questa famiglia.

## V. Il principio della costruttività condizionata

Il meccanismo dell'ascesa gödeliana e le sue istanze nei teoremi diagonali dimostrano che certi incontri con l'impossibilità producono strutture nuove. Ma prima di estendere la tesi agli altri domini, è necessario qualificarla con precisione: non ogni incontro con un'impossibilità è costruttivo. La costruttività è condizionata — dipende da proprietà specifiche del sistema che incontra

l'impossibilità — e le condizioni della costruttività sono parte essenziale della teoria.

Si chiama principio della costruttività condizionata (PCC) l'insieme delle condizioni necessarie e congiunte per cui il surplus generativo  $\Gamma(I, S)$  di un'impossibilità strutturale  $I$  rispetto a un sistema  $S$  viene convertito in strutture nuove. Le condizioni sono quattro.

La prima condizione è il riconoscimento strutturale: il sistema  $S$  deve essere sufficientemente ricco da distinguere l'impossibilità strutturale  $I$  da un fallimento contingente delle proprie risorse attuali. Un sistema che interpreta un'impossibilità come un problema tecnico temporaneo — qualcosa che risorse maggiori o algoritmi migliori risolverebbero — non inizia il processo di generazione di strutture nuove perché non riconosce la necessità dell'estensione. Questa condizione è più sottile di quanto appaia. Non si tratta semplicemente di avere abbastanza conoscenza tecnica: si tratta di avere la struttura cognitiva per distinguere tra due tipi di limitazione che, dall'interno del sistema, possono sembrare identici. Un sistema formale  $T$  non sa, dall'interno, se  $G_T$  è indecidibile perché è troppo debole (problema contingente) o perché  $G_T$  è genuinamente indecidibile in qualsiasi estensione di tipo  $T$  (problema strutturale). L'inferenza che si tratta del secondo tipo richiede risorse metalinguistiche — richiede il punto di vista semantico esterno che il teorema di Tarski dice essere irraggiungibile dall'interno. La condizione di riconoscimento strutturale è quindi essa stessa soggetta all'impossibilità di chiusura semantica: un sistema non può, dall'interno, determinare con certezza assoluta se sta incontrando un'impossibilità strutturale. Può soltanto avere elementi di evidenza sufficienti per la distinzione, e questi elementi sono sempre, in qualche misura, provenienti da un punto di vista parzialmente esterno — dalla matematica della teoria, dall'interpretazione in modelli, dall'analisi della struttura degli argomenti.

Il caso di Wittgenstein è l'esempio più istruttivo di mancato riconoscimento strutturale. Nelle *Bemerkungen über die Grundlagen der Mathematik* (pubblicate postume nel 1956), Wittgenstein tratta il teorema di Gödel come un gioco di parole — una costruzione linguistica interessante ma priva di significato matematico genuino. La sentenza  $G_{\{PA\}}$ , per Wittgenstein, non "dice" nulla su se stessa: è una sequenza di simboli con certe regole di manipolazione, e

l'interpretazione che "dice di essere indimostrabile" è un'aggiunta filosofica esterna che non cambia la struttura matematica del sistema. Questa lettura viola la prima condizione: Wittgenstein non riconosce  $G_{\{PA\}}$  come un'impossibilità strutturale di PA ma come un artefatto della codifica, qualcosa di contingente nella scelta della gödelizzazione. Il risultato è che, per la filosofia della matematica di Wittgenstein, il teorema di Gödel non genera alcuna struttura nuova — è neutralizzato prima di poter essere costruttivo.

La seconda condizione è la persistenza operativa: il sistema S deve rimanere operativo attraverso l'incontro con I abbastanza a lungo da avviare la costruzione di S'. Questa condizione esclude i sistemi che si dissolvono nell'incontro — che subiscono un collasso delle proprie strutture prima che la distinzione generata dall'impossibilità possa essere integrata. La condizione sembra ovvia, ma nei sistemi biologici e computazionali è tutt'altro che garantita. Il catastrophic forgetting nei sistemi di apprendimento artificiale (capitolo X di questo ciclo) è precisamente il fallimento di questa condizione: quando un sistema di reti neurali viene addestrato su un nuovo compito che interferisce con le strutture apprese in precedenza, può perdere le strutture precedenti prima che la nuova struttura sia consolidata. Non si genera un sistema S' più ricco che integra entrambi i compiti: si ottiene semplicemente la distruzione del vecchio a vantaggio del nuovo. La mancanza di persistenza trasforma l'impossibilità in evento distruttivo invece che costruttivo.

La terza condizione è la differenziazione genuina: l'impossibilità I deve produrre in S una distinzione che non era articolabile nella struttura di S prima dell'incontro. Questa condizione è logicamente la più delicata. Non basta che il sistema riconosca l'impossibilità: è necessario che l'impossibilità articoli una separazione tra due aspetti del dominio che prima erano confusi o non distinti. Nel caso dell'ascesa gödeliana, la distinzione è tra verità e dimostrabilità:  $G_{\{PA\}}$  è vera (nella semantica standard) ma indimostrabile (nella struttura deduttiva di PA), e questa distinzione — che prima della sentenza gödeliana non aveva un referente formale preciso — è esattamente ciò che genera la necessità di distinguere tra semantica e sintassi in modo tecnico, di costruire la nozione di modello come separata da quella di derivazione, di sviluppare la teoria dei modelli come campo autonomo. La distinzione tra verità e dimostrabilità non era "già saputa" e semplicemente non formalizzata: non aveva nemmeno la forma di un'intuizione

chiara prima di  $G_{\{PA\}}$ . L'impossibilità ha prodotto la distinzione, non soltanto formalizzato una distinzione preesistente.

La quarta condizione è l'integrabilità:  $S$ , o una sua estensione minimale compatibile con la struttura di  $S$ , deve avere la struttura sufficiente a integrare la distinzione generata da  $I$  in un sistema  $S'$  che risolve problemi che  $S$  non poteva risolvere. Questa è la condizione della costruttività in atto — la condizione che separa il surplus generativo in potenza dal surplus generativo effettivamente convertito in struttura. La condizione può non essere soddisfatta immediatamente: il paradosso di Zenone ha avuto surplus generativo positivo per più di duemila anni prima che la struttura matematica dell'ambiente intellettuale raggiungesse la maturità necessaria per soddisfare la condizione di integrabilità. Il calcolo infinitesimale di Newton e Leibniz (1665-1684 circa) è il sistema  $S'$  che integra la distinzione prodotta dal paradosso di Zenone — la distinzione tra somme infinite divergenti e somme infinite convergenti, tra infiniti attuali e processi limite — e che risolve il problema dell'accelerazione e del moto che Zenone aveva reso acuto. Ma questa integrazione ha richiesto lo sviluppo della nozione di limite, di funzione, di derivata e integrale — strutture che non erano disponibili nella matematica greca, medievale o rinascimentale. La costruttività del paradosso di Zenone era condizionata allo sviluppo del contesto matematico ambientale.

Le quattro condizioni sono necessarie e congiunte: la mancanza di una sola è sufficiente a impedire la costruttività. Questo è un punto importante perché significa che la paradossologia costruttiva non afferma che le impossibilità strutturali siano sempre e ovunque costruttive — afferma che lo sono quando le condizioni sono soddisfatte, e che la forma precisa della costruttività è determinata dalla forma precisa dell'impossibilità. La teoria non è ottimista in modo indiscriminato: è precisa sulle condizioni della propria applicabilità.

## **VI. La costruttività nella Famiglia II: il confine dimensionale come mappa architettonica**

Le impossibilità dimensionali — la Famiglia II — hanno un meccanismo di costruttività diverso da quello delle diagonali, e la differenza è istruttiva. Le impossibilità diagonali producono strutture nuove attraverso il salto a un metalinguaggio superiore — la struttura nuova è categorialmente diversa dalla

struttura precedente, incompatibile con essa in certi aspetti, necessariamente superiore in potenza espressiva. Le impossibilità dimensionali producono strutture nuove attraverso la precisazione della frontiera — la struttura nuova è la caratterizzazione esatta del confine, e questa caratterizzazione rende possibile l'ottimizzazione sul confine stesso.

Il teorema di Johnson-Lindenstrauss (1984) — dimostrato nella sua completezza nell'Appendice B — afferma che  $m$  punti in  $\mathbb{R}^D$  possono essere proiettati in uno spazio di dimensione  $n = O(\log m / \epsilon^2)$  preservando tutte le distanze con distorsione al più  $\epsilon$ . Il converse — dimostrato da Alon (2003) nella forma combinatoriale e da Jayram-Woodruff (2013) nella forma della complessità della comunicazione — afferma che questa dimensione è ottimale: nessuna proiezione in dimensione inferiore a  $\Omega(\log m / \epsilon^2)$  può preservare le distanze con distorsione  $\epsilon$ . Il confine  $n \cdot \epsilon^2 = \Theta(\log m)$  è quindi sharp — un lower bound e un upper bound che coincidono a meno di costanti.

Prima del teorema di JL e del suo converse, le proiezioni di dimensione ridotta esistevano come tecniche (l'analisi delle componenti principali era nota da Pearson 1901), ma erano costruite sulla base di criteri euristici — si proiettava nelle direzioni di massima varianza, o si usavano altre regole di buon senso. Non si sapeva se esistesse una teoria della proiezione ottimale, né quale fosse il prezzo che ogni proiezione deve necessariamente pagare. L'impossibilità dimensionale — il fatto che qualsiasi proiezione da  $\mathbb{R}^m$  in  $\mathbb{R}^n$  con  $m > n$  è necessariamente non-isometrica per qualche coppia di punti — non aveva ancora prodotto la sua distinzione strutturale.

La distinzione prodotta dalla caratterizzazione precisa del confine è la seguente: esiste un regime di dimensionalità ( $n \geq \Theta(\log m / \epsilon^2)$ ) in cui la fedeltà è garantita, e un regime ( $n < \Theta(\log m / \epsilon^2)$ ) in cui non lo è. Questa distinzione non è semplicemente "sopra il confine va bene, sotto va male": è la specificazione di una frontiera quantitativa che separa i sistemi rappresentativi ottimali (quelli che si trovano sul confine) da quelli sub-ottimali (quelli che si trovano sotto) e da quelli ridondanti (quelli che si trovano sopra). I sistemi ottimali — quelli che usano esattamente la dimensione necessaria e sufficiente — sono i sistemi che abitano il confine.

Le strutture prodotte dalla conoscenza precisa di questo confine sono numerose e di notevole rilevanza tecnica. La compressed sensing (Candès, Romberg, Tao 2006; Donoho 2006) è la teoria che mostra come i segnali sparsi — segnali che hanno soltanto  $k$  componenti non nulle in qualche base, con  $k \ll m$  — possono essere recuperati esattamente da  $n = O(k \log(m/k))$  misurazioni invece delle  $m$  misurazioni che il campionamento di Nyquist richiederebbe. La struttura formale usa la Restricted Isometry Property (RIP): una matrice di misurazione  $A \in \mathbb{R}^{n \times m}$  soddisfa la RIP di ordine  $k$  con costante  $\delta$  se:

$$(1 - \delta) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta) \|x\|_2^2$$

per ogni vettore  $k$ -sparso  $x$ . Quando  $\delta_{\{2k\}} < \sqrt{2} - 1$ , il recupero esatto è garantito attraverso la minimizzazione della norma  $\ell_1$ . La RIP è la versione quasi-isometrica di JL per vettori sparsi — è la struttura che rende operativa la conoscenza del confine dimensionale nel contesto specifico dei segnali sparsi. La compressed sensing non avrebbe potuto essere costruita senza la teoria del confine dimensionale: la RIP è definibile soltanto se si ha la nozione di "quasi-isometria" che JL articola.

La costruttività delle impossibilità dimensionali non ha la stessa evidenza formale dell'ascesa gödeliana — non si può scrivere una sequenza di sistemi  $T_0 \subsetneq T_1 \subsetneq T_2 \subsetneq \dots$  con ordinali di prova strettamente crescenti. Ma ha un'evidenza storica e teorica che non può essere ignorata: la compressed sensing, le random projections algoritmiche, la teoria dei manifolds per l'apprendimento automatico sono tutte strutture che non erano concepibili senza la comprensione esatta del confine che JL e il suo converse hanno articolato. Il surplus generativo  $\Gamma(\text{Confine\_JL}, \text{Matematica\_pre-JL})$  — misurato dalla complessità di Kolmogorov della teoria della compressed sensing condizionata alla conoscenza dell'analisi funzionale pre-JL — è positivo e non banale. La costruttività è reale, anche se la sua misura formale è meno precisa di quella delle diagonalità.

Nel dominio biologico, la costruttività del binding problem (l'impossibilità dimensionale più rilevante per la neurobiologia cognitiva) ha una storia evolutiva. Il sistema nervoso biologico ha incontrato l'impossibilità del binding esatto attraverso il processo evolutivo — non attraverso una scoperta matematica — e la risposta dell'evoluzione alla forma precisa di questa impossibilità è la codifica sparsa. Il fatto che le cellule piramidali corticali siano tipicamente attive soltanto

nel 2-5% delle situazioni (invece di essere attive sempre o in proporzione uniforme) non è un'inefficienza evolutiva: è la soluzione ottimale al confine dimensionale che il binding problem impone. La sparsità  $\varphi$  mantiene l'interferenza di binding a livelli tollerabili: l'ampiezza del rumore di recupero è proporzionale a  $\varphi \cdot m$  (dove  $m$  è il numero di feature e  $\varphi$  la sparsità), e per  $\varphi \approx 0.02$  e  $m \approx 10^4$  l'interferenza è dell'ordine di 200 — gestibile per un sistema con  $n \approx 10^6$  neuroni. Un sistema con attività densa ( $\varphi \approx 0.5$ ) avrebbe un'interferenza di 5000 volte maggiore, rendendo impossibile il binding di oggetti multipli nella stessa scena.

L'evoluzione non "conosce" il teorema di JL, ma ha trovato — attraverso miliardi di anni di selezione su sistemi che dovevano distinguere oggetti in scene affollate — esattamente la soluzione che il teorema di JL prescrive come ottimale per i parametri biologici disponibili. Questa convergenza tra la soluzione evolutiva e la soluzione matematicamente ottimale è uno dei fenomeni più suggestivi della biologia teorica, e si capisce soltanto se si comprende la struttura dell'impossibilità dimensionale che l'evoluzione stava risolvendo.

## **VII. La costruttività nella Famiglia III: l'intrattabilità come mappa del regime di approssimazione**

Le impossibilità di complessità computazionale hanno una costruttività che opera attraverso un meccanismo diverso dai due precedenti. Non producono metalinguaggi superiori, né architetture rappresentative ottimali: producono la teorizzazione del regime di approssimazione come oggetto matematico autonomo — la scoperta che esistono gradi di difficoltà nel paecapitolo dei problemi, che questi gradi hanno struttura, e che la struttura dei gradi ha conseguenze normative per il design dei sistemi che operano in quei regimi.

La NP-hardness del problema del commesso viaggiatore (TSP) — il problema di trovare il ciclo hamiltoniano di peso minimo in un grafo completo con pesi sugli archi — è nota dal 1972 (Karp). Ma la dimostrazione che il TSP è NP-hard non ha prodotto il nulla: ha prodotto la teoria delle approssimazioni algoritmiche come campo autonomo. La distinzione fondamentale che l'impossibilità di complessità ha generato è tra "risolvere il problema esattamente" (impossibile in tempo polinomiale se  $P \neq NP$ ) e "trovare una soluzione entro un fattore costante dall'ottimale in tempo polinomiale" (possibile, con garanzie precisamente

quantificabili). Questa distinzione non era articolata prima della NP-hardness: i matematici cercavano algoritmi esatti, e la mancanza di algoritmi esatti efficienti era interpretata come fallimento tecnico. La NP-hardness ha convertito questo fallimento in un fatto strutturale, e il fatto strutturale ha generato la domanda: quale è il migliore rapporto di approssimazione raggiungibile in tempo polinomiale?

L'algoritmo di Christofides-Serdyukov (1976), per il TSP metrico (dove le distanze soddisfano la disuguaglianza triangolare), garantisce un tour di costo al più  $3/2 \cdot \text{OPT}$  in tempo polinomiale. Questo non è un risultato di resa di fronte all'intrattabilità: è la caratterizzazione del confine di approssimabilità dalla parte del possibile. La garanzia  $3/2$  non è arbitraria: è l'unica garanzia che si può ottenere combinando matching di peso minimo, spanning tree di peso minimo, e la struttura del tour attraverso la disuguaglianza triangolare. La forma precisa dell'algoritmo di Christofides è determinata dalla forma precisa del confine tra ciò che è ottimizzabile esattamente e ciò che non lo è.

Il PCP theorem (Arora-Safra 1998, Arora et al. 1998) porta questa struttura al livello più profondo. La caratterizzazione  $\text{NP} = \text{PCP}(O(\log n), O(1))$  — ogni problema in NP ha prove verificabili con  $O(\log n)$  bit aleatori e  $O(1)$  query al testo della prova — implica, attraverso la costruzione esplicita dei gap di approssimabilità, che certi problemi non sono approssimabili al di là di una soglia precisa in tempo polinomiale (se  $P \neq \text{NP}$ ). Per MAX-3SAT, Håstad (2001) dimostra che il gap è esattamente  $7/8$ : nessun algoritmo polinomiale può garantire più di  $7/8$  dell'ottimo su tutte le istanze, e l'algoritmo casuale banale (assegna i valori di verità uniformemente a caso) raggiunge esattamente  $7/8$  in attesa. Il confine di approssimabilità per MAX-3SAT è quindi sharp — non c'è nulla da guadagnare con algoritmi più sofisticati, nell'estremo peggiore.

L'analogia con il confine di JL è strutturale: in entrambi i casi, l'impossibilità definisce una frontiera con un upper bound e un lower bound che coincidono, e la conoscenza di questa frontiera rende possibile l'ottimizzazione sulla frontiera stessa. Il surplus generativo della NP-hardness — misurato dalla complessità di Kolmogorov dell'intera teoria delle approssimazioni algoritmiche condizionata alla matematica discreta pre-1970 — è molto grande.

Un avvertimento metodologico è però necessario qui, e non va omissso. La costruttività delle impossibilità di complessità dipende dalla verità della congettura  $P \neq NP$ . Se  $P = NP$ , il confine tra "risolvibile esattamente in tempo polinomiale" e "risolvibile soltanto approssimativamente in tempo polinomiale" non esiste: i due regimi coincidono. In quel caso, molti dei gap di approssimabilità scomparirebbero, e la teoria dell'approssimazione cambierebbe radicalmente di fisionomia. La costruttività di questa famiglia è quindi epistemologicamente più fragile delle precedenti: è condizionata alla verità di una congettura. La comunità matematica ritiene quasi certamente  $P \neq NP$  — su questo la letteratura tecnica e i risultati di Baker-Gill-Solovay (1975) convergono — ma una quasi-cerchezza non è una dimostrazione. La paradossologia costruttiva per la Famiglia III ha un margine di incertezza che le Famiglie I e IV non hanno.

### **VIII. La costruttività nella Famiglia IV: la gerarchia tarskiana come struttura necessaria**

Le impossibilità di chiusura semantica — la Famiglia IV — hanno la costruttività più direttamente formalizzabile dopo quella della Famiglia I, perché il teorema di Tarski è dimostrato con gli stessi strumenti (il lemma della diagonalizzazione) che fondano l'ascesa gödeliana.

Il teorema di Tarski afferma che per ogni linguaggio formale  $L_0$  consistente e sufficientemente ricco da contenere l'aritmetica elementare, non esiste un predicato  $Tr_0(x)$  di  $L_0$  che soddisfi lo schema T — la condizione che  $Tr_0(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$  per ogni sentenza  $\varphi$  di  $L_0$ . La dimostrazione usa il lemma della diagonalizzazione per costruire  $G$  tale che  $L_0 \vdash G \leftrightarrow \neg Tr_0(\ulcorner G \urcorner)$  e ricavare la contraddizione. Il predicato di verità non è definibile dall'interno.

Ma cosa produce questa impossibilità? Produce necessariamente  $L_1$  — il metalinguaggio che può definire il predicato di verità per  $L_0$ . La costruzione tarskiana della soddisfazione (1933) è esplicitamente la costruzione di  $Tr_0$  in  $L_1$ : la definizione ricorsiva di "la sentenza  $\varphi$  è vera nel modello  $M$ " procede sui livelli della struttura logica di  $\varphi$  — vero per le formule atomiche (dipende dall'interpretazione dei simboli non logici), esteso alle negazioni, alle congiunzioni, ai quantificatori secondo regole ricorsive esplicite. La costruzione è effettiva: produce, in  $L_1$ , un predicato che cattura esattamente la semantica di  $L_0$ .

L'impossibilità di definire  $Tr_0$  in  $L_0$  ha generato, necessariamente,  $L_1$  come il livello minimo in cui  $Tr_0$  diventa definibile.

La sequenza si itera.  $L_1$  non può definire il proprio predicato di verità  $Tr_1$  dall'interno — lo stesso argomento di Tarski si applica a  $L_1$ . Ma  $Tr_1$  è definibile in  $L_2$ . La gerarchia tarskiana:

$$L_0 \subsetneq L_1 \subsetneq L_2 \subsetneq \dots$$

con  $L_n$  in grado di definire il predicato di verità per  $L_{n-1}$ , è la struttura prodotta dall'iterazione dell'impossibilità di chiusura semantica. La proprietà formale della gerarchia è:

$$\text{Exp}(L_n) \subsetneq \text{Exp}(L_{n+1}) \text{ per ogni } n$$

dove  $\text{Exp}(L_n)$  è la classe delle verità esprimibili in  $L_n$ . La differenza  $\text{Exp}(L_{n+1}) \setminus \text{Exp}(L_n)$  contiene almeno il predicato di verità di  $L_n$  — un insieme di verità che  $L_n$  non può catturare ma che  $L_{n+1}$  può. Il surplus generativo di ciascun pascapitolo è  $K(\text{Exp}(L_{n+1}) \setminus \text{Exp}(L_n) \mid L_n) > 0$ , positivo per ogni  $n$ .

La gerarchia tarskiana non è soltanto una costruzione tecnica per evitare il paradosso del mentitore: è la struttura semantica di qualsiasi sistema di conoscenza che voglia parlare della propria verità. Ogni volta che un sistema cognitivo tenta di verificare le proprie asserzioni — di chiedersi se le proprie rappresentazioni del mondo corrispondono al mondo — sta tentando di costruire il predicato di verità per il proprio linguaggio interno. Il teorema di Tarski dice che questa operazione richiede un punto di vista esterno al sistema stesso: non ci si può fidare soltanto della coerenza interna, perché la coerenza interna non implica la verità nella semantica esterna.

Nel dominio biologico, la separazione wanting/liking di Berridge — la dissociazione tra il sistema dopaminergico (wanting, la direzionalità dell'azione) e il sistema oppioide (liking, la valorizzazione del risultato) — ha una struttura che riflette esattamente questa impossibilità. Il sistema dopaminergico implementa la "sintassi" del valore: codifica le relazioni predittive tra segnali e risultati, aggiorna i pesi in proporzione all'errore di previsione, dirige il comportamento verso stati futuri stimati come preferibili. Il sistema oppioide implementa la "semantica" del valore: codifica il significato intrinseco dell'esperienza, il fatto che certe sensazioni

siano piacevoli invece che soltanto predette come piacevoli. La dissociazione tra i due sistemi — il fatto che li si possa separare farmacologicamente, che esistano pazienti con wanting senza liking e liking senza wanting — è la struttura biologica che corrisponde all'impossibilità di chiudere la semantica all'interno della sintassi. Il cervello ha costruito due sistemi separati operanti a livelli diversi della gerarchia biologica del valore — non per inefficienza architettonica, ma perché è impossibile realizzare entrambe le funzioni in un sistema unico senza interferenze strutturali.

## **IX. La costruttività nella Famiglia V: l'irreversibilità come condizione dell'identità**

Le impossibilità di irreversibilità — la Famiglia V — hanno la costruttività più controintuitiva, perché l'irreversibilità è tipicamente vissuta come perdita. Una struttura che non può essere invertita è una struttura che ha perso delle opzioni — un sistema che ha subito un processo irreversibile è meno plastico, meno adattabile, meno capace di tornare a stati precedenti. Come può questa perdita essere generativa?

La risposta richiede di distinguere tra due livelli di analisi. Al primo livello — il livello del singolo processo irreversibile — la perdita è reale. Il ratchet di Müller (capitolo VI di questo ciclo) accumula mutazioni deleterie in modo irreversibile, e ogni scatto del ratchet è un deterioramento del genoma. La chiusura di un periodo critico è una perdita di plasticità — certe strutture di apprendimento non sono più accessibili. La deriva epigenetica è degradazione dell'informazione regolatoria. Questi non sono fenomeni costruttivi al livello del singolo processo.

Al secondo livello — il livello della pressione selettiva o evolutiva — l'irreversibilità strutturale è la condizione che rende necessarie strutture che la reversibilità non richiederebbe e che non potrebbe produrre. La costruttività opera a questo secondo livello.

Il ratchet di Müller opera nelle popolazioni asessuate con velocità  $v$  che dipende dai parametri biologici della popolazione (tasso di mutazione  $\mu$ , coefficiente di selezione  $s$ , dimensione effettiva  $N$ ). Per  $v$  sufficientemente alta, il costo metabolico del twofold del sesso (le femmine sessuate producono la metà dei figli fertili rispetto alle femmine asessuate, perché metà dei figli sono maschi) è

superato dal vantaggio della ricombinazione. La formula dell'equilibrio — derivata da Maynard Smith (1978) e Kondrashov (1988) in varianti diverse — afferma che il sesso è selezionato quando la velocità del ratchet supera una soglia che dipende dai parametri della popolazione. La struttura critica è:

Il sesso è vantaggioso  $\leftrightarrow v > v_{\text{critico}}(\mu, s, N)$

Questa è una relazione quantitativa tra la velocità dell'impossibilità ( $v$ ) e la struttura della risposta che l'impossibilità genera (il sesso). La forma della risposta — la ricombinazione genetica come meccanismo per ricostituire genomi a basso numero di mutazioni attraverso la combinazione di linee genetiche distinte — è determinata dalla forma dell'impossibilità: il ratchet è irreversibile perché l'informazione sul genoma originale è perduta quando la classe zero si estingue, e la ricombinazione è esattamente il meccanismo che ricostruisce approssimazioni al genoma originale combinando parti di genomi diversi che hanno preservato parti distinte del genoma ancestrale. Il sesso non è una risposta generica all'impossibilità — è la risposta determinata dalla struttura specifica dell'impossibilità.

$K(\text{Sesso} \mid \text{Asessuatità})$  è positivo nel senso tecnico: la riproduzione sessuale è una struttura biologica genuinamente più complessa della riproduzione asessuata (richiede meccanismi di compatibilità sessuale, di meiosi, di fecondazione, di sviluppo dell'embrione da due genomi), e questa complessità aggiuntiva è il costo dell'informazione genetica guadagnata. L'impossibilità del ratchet ha prodotto strutture che non sarebbero emerse in popolazioni asessuate al sicuro dal ratchet (popolazioni piccole e stabili).

La chiusura del periodo critico ha un meccanismo di costruttività ancora più diretto. Il periodo critico non è una struttura indipendente che poi si chiude: è esso stesso definito dalla chiusura. Il sistema nervoso in sviluppo attraversa una fase di alta plasticità in cui le connessioni sono formate, rinforzate, e potate in risposta all'esperienza — poi la plasticità diminuisce e le strutture si stabilizzano. La stabilizzazione è l'impossibilità di irreversibilità: i circuiti consolidati resistono alla modifica, le sinapsi potate non vengono ristabilite, certi apprendimenti diventano molto difficili o impossibili.

La struttura cognitiva prodotta da questo processo — la competenza adulta — ha proprietà che il sistema plastico infantile non ha e non può avere: velocità di elaborazione, affidabilità nel tempo, robustezza alle perturbazioni, capacità di

operare in modo automatico senza costo attentivo. Queste proprietà emergono precisamente dalla riduzione della plasticità: un sistema interamente plastico non può sviluppare la velocità e l'automaticità della competenza adulta perché ogni operazione richiederebbe il ricalcolo delle connessioni in risposta al contesto. La plasticità è il prezzo dell'apprendimento; la sua riduzione è il guadagno della competenza. Non è possibile avere entrambi simultaneamente — questa è l'impossibilità — e la struttura della competenza adulta è il prodotto necessario dell'impossibilità di mantenere la plasticità indefinitamente.

Si noti la struttura formale comune tra il sesso e la competenza adulta come risposte alle impossibilità della Famiglia V: in entrambi i casi, la risposta costruttiva produce strutture che hanno proprietà inaccessibili al sistema prima dell'impossibilità (la ricombinazione genetica; la velocità e l'affidabilità della competenza), e queste proprietà sono determinate dalla forma specifica dell'impossibilità (l'irreversibilità dell'accumulo di mutazioni; l'irreversibilità della potatura sinaptica). La costruttività non è generica — è specifica alla forma del confine.

## **X. Il salto emergente: la biforcazione come struttura della generatività**

Le strutture prodotte dall'incontro con le impossibilità strutturali — i metalinguaggi gödeliani, le architetture di codifica sparsa, gli algoritmi di approssimazione, la gerarchia tarskiana, il sesso e la competenza adulta — non emergono attraverso processi gradualisti e continui. Emergono attraverso salti: discontinuità qualitative in cui la struttura del sistema cambia categorialmente in risposta al raggiungimento di una soglia critica. Questo fenomeno — che chiamo il salto emergente — ha una struttura formale precisa nella teoria matematica delle biforcazioni, e comprenderlo è necessario per capire perché le strutture prodotte dall'impossibilità sono genuinamente nuove nel senso informativo.

La teoria delle biforcazioni (le cui origini risalgono a Poincaré e che è stata sistematizzata da Andronov, Hopf, e Thom nel Novecento) studia i modi in cui la struttura qualitativa di un sistema dinamico cambia al variare di un parametro. Sia  $\dot{x} = F(x, \lambda)$  un sistema dinamico parametrizzato da  $\lambda \in \mathbb{R}$ . Per un dato  $\lambda$ , il sistema

ha un insieme di punti fissi, cicli limite, e attrattori. Al variare di  $\lambda$ , questi oggetti cambiano qualitativamente in certi punti critici  $\lambda^*$  — le biforcazioni.

Tre tipi di biforcazione sono rilevanti per la paradossologia costruttiva.

La biforcazione nodo-sella: per  $\lambda < \lambda^*$  non esistono punti fissi in una regione dello spazio degli stati; per  $\lambda = \lambda^*$  appare un unico punto fisso (il punto di sella); per  $\lambda > \lambda^*$  esistono due punti fissi, uno stabile e uno instabile. Il sistema "guadagna" un attrattore che prima non esisteva — non attraverso un processo graduale, ma attraverso una creazione discontinua al valore critico  $\lambda^*$ . Questa struttura è quella del periodo critico nello sviluppo neurale: per una certa finestra temporale ( $\lambda < \lambda^*$ ), la struttura del sistema nervoso è in uno stato di alta plasticità senza un attrattore stabile consolidato; al raggiungimento della soglia critica ( $\lambda = \lambda^*$ ), i circuiti si cristallizzano in un attrattore stabile che prima non esisteva.

La biforcazione transcritica: per  $\lambda < \lambda^*$  il sistema ha due punti fissi, uno stabile e uno instabile; a  $\lambda = \lambda^*$  si scambiano la stabilità; per  $\lambda > \lambda^*$  quello che era stabile diventa instabile e viceversa. Il sistema attraversa una transizione di regime in cui l'equilibrio precedente viene destabilizzato e uno nuovo prende il suo posto. Questa è la struttura della rivoluzione scientifica di Kuhn: il paradigma dominante (l'attrattore stabile per  $\lambda < \lambda^*$ ) si destabilizza al crescere delle anomalie (il parametro  $\lambda$ ), e il nuovo paradigma (il vecchio equilibrio instabile) diventa il nuovo stato stabile per  $\lambda > \lambda^*$ . La transizione non è graduale: è la destabilizzazione dell'attrattore precedente e la stabilizzazione del successivo.

La biforcazione di Hopf: per  $\lambda < \lambda^*$  il sistema ha un punto fisso stabile; a  $\lambda = \lambda^*$  il punto fisso perde la stabilità e per  $\lambda > \lambda^*$  compare un ciclo limite stabile. Il sistema passa da un comportamento stazionario a un comportamento oscillatorio. Questa è la struttura dell'insight cognitivo come la neurobiologia lo descrive: il sistema cognitivo in uno stato di "incubazione" (il punto fisso stabile del vecchio problema) attraversa una destabilizzazione al raggiungimento di certe condizioni (il parametro critico  $\lambda^*$ ) e emerge in un regime di "esplorazione oscillatoria" (il ciclo limite) che poi si stabilizza in una nuova soluzione (un nuovo punto fisso stabile dopo la biforcazione).

Il punto teoricamente cruciale è che in tutti e tre i tipi di biforcazione la struttura che emerge per  $\lambda > \lambda^*$  — il nuovo attrattore, il nuovo regime stabile, il ciclo limite — non è visibile come struttura dallo stato del sistema per  $\lambda < \lambda^*$ . Un osservatore

che conosce il sistema soltanto per  $\lambda < \lambda^*$  non può dedurre la struttura che apparirà per  $\lambda > \lambda^*$ : quella struttura emerge nella biforcazione, non prima. Formalmente:  $K(\text{Struttura\_post-biforcazione} \mid \text{Struttura\_pre-biforcazione}) > 0$  — la struttura successiva porta informazione genuinamente nuova rispetto alla struttura precedente.

Questa è la formalizzazione esatta del surplus generativo per i sistemi biforcativi. Le strutture prodotte dall'impossibilità —  $T_1$  rispetto a  $T_0$  nell'ascesa gödeliana, la compressed sensing rispetto alla matematica pre-JL, il sesso rispetto all'asessualità — sono tutte strutture post-biforcazione nel senso matematico preciso: non erano prevedibili dalla struttura pre-biforcazione perché emergono nel pascapitolo attraverso il confine, non prima. La non-prevedibilità non è un difetto della previsione: è la misura dell'informazione genuinamente nuova che il confine produce.

Si può ora formulare la paradossologia costruttiva in modo più preciso di quanto la sezione I abbia fatto. Le impossibilità strutturali del PAI sono le strutture che definiscono i valori critici  $\lambda^*$  nel paecapitolo cognitivo — i punti in cui la struttura del sistema attraversa una biforcazione. I sistemi che si trovano in prossimità di questi valori critici sono i sistemi al confine nel senso topologico: sono nella regione di massima sensibilità alle perturbazioni, dove piccoli cambiamenti nella struttura producono trasformazioni qualitative nell'insieme degli attrattori. Il confine è il luogo di massima informazione potenziale — il luogo da cui la struttura nuova sta per emergere.

## **XI. Il limite della paradossologia costruttiva: quando l'impossibilità è distruttiva**

Una teoria intellectualmente onesta deve specificare non soltanto quando la propria tesi vale, ma quando non vale. La paradossologia costruttiva afferma che le impossibilità strutturali sono costruttive sotto le condizioni del PCC. È necessario sviluppare con uguale precisione i casi in cui queste condizioni non sono soddisfatte, perché questi casi rivelano la struttura della teoria da una prospettiva che i casi positivi non mostrano.

Il fallimento della costruttività per mancanza di riconoscimento strutturale — la prima condizione — ha il caso paradigmatico nella risposta di Frege al paradosso di Russell. Frege, nella famosa appendice al secondo volume dei *Grundgesetze* scritta mentre il libro era già in stampa, tenta di salvare il sistema attraverso una modifica della Legge Base V: invece di affermare che ogni proprietà  $F$  ha un'estensione  $\{x : F(x)\}$ , afferma che per ogni proprietà  $F$  la classe  $\{x : F(x)\}$  esiste come oggetto ma coincide con un oggetto "segnalato" speciale quando  $F$  non è applicabile all'estensione stessa. Questa modifica — nota come Legge Base V' — è inconsistente anch'essa, come Lesniewski dimostrò nel 1939. Frege stava cercando una soluzione locale — una modifica minimale che preservasse la struttura fondamentale del sistema — invece di riconoscere che il problema richiedeva una ristrutturazione categoriale. La mancanza di riconoscimento strutturale ha impedito la costruttività: Frege non ha costruito la teoria dei tipi né ZF, è rimasto nella logica che aveva costruito, tentando di tappare un buco strutturale con una modifica sintattica.

Il fallimento della costruttività per mancanza di persistenza operativa — la seconda condizione — è esemplificato nei sistemi cognitivi e biologici che si dissolvono nell'incontro con l'impossibilità invece di rispondervi. La malattia di Alzheimer è, dal punto di vista funzionale, un fallimento della persistenza: il sistema cognitivo incontra l'impossibilità crescente di mantenere la coerenza tra le strutture semantiche (i ricordi, i concetti, le relazioni) e le strutture sintattiche del cervello (le connessioni neurali, le sinapsi), ma invece di generare strutture nuove che abitino il confine in modo produttivo, si dissolve progressivamente — le strutture esistenti si degradano senza essere sostituite da strutture nuove. Il risultato è la perdita di funzione cognitiva, non la generazione di nuova struttura cognitiva.

Il fallimento della costruttività per mancanza di differenziazione genuina — la terza condizione — è il caso dei sistemi che incontrano l'impossibilità ma non producono la distinzione che l'impossibilità richiederebbe. Un esempio è la tradizione medioevale nel trattare il problema degli universali (se i concetti generali come "umanità" o "bianchezza" esistono come oggetti reali o siano soltanto nomi). Il problema degli universali è, dal punto di vista formale, una versione del confine sintassi-semantica: la domanda è se i predicati del linguaggio abbiano referenti nel dominio semantico (realismo) o siano soltanto strumenti sintattici

(nominalismo). Ma la disputa scolastica ha prodotto, per diversi secoli, posizioni che si opponevano senza articolare una distinzione nuova — il dibattito tra realisti e nominalisti riproduceva le stesse strutture argomentative senza che nessuna delle parti producesse una struttura concettuale che non fosse già presente nella formulazione originale del problema. La distinzione genuina — quella che la teoria dei tipi o la teoria dei modelli articolerà nel Novecento — non è emersa dalla disputa medioevale, non per mancanza di intelligenza dei partecipanti, ma per mancanza degli strumenti formali necessari a rendere la distinzione precisa.

Il fallimento della costruttività per mancanza di integrabilità — la quarta condizione — è il caso più comune nella storia della scienza: il caso del risultato che anticipa il proprio contesto. Boltzmann, nella seconda metà dell'Ottocento, sviluppa la meccanica statistica — la derivazione delle leggi termodinamiche a partire dalle equazioni del moto delle molecole attraverso l'analisi statistica. Il programma incontra opposizione durissima da fisici come Mach e Ostwald, non perché il programma sia sbagliato, ma perché le strutture matematiche necessarie per renderlo preciso — la misura di Lebesgue, la teoria degli spazi di probabilità, l'ergodi teoria — non esistevano ancora nella forma necessaria. Boltzmann stava producendo il surplus generativo di una distinzione reale (tra la meccanica microscopica reversibile e la termodinamica macroscopica irreversibile), ma la condizione di integrabilità non era soddisfatta. Il risultato fu l'isolamento intellettuale e infine il suicidio di Boltzmann nel 1906. La costruttività era in potenza — è stata integrata nel Novecento da Gibbs, Einstein, e poi dalla meccanica statistica moderna — ma non era in atto nel momento dell'incontro con l'impossibilità.

## **XII. Verso la sintesi: il territorio come struttura biforcata**

La paradossologia costruttiva, sviluppata nei termini tecnici delle sezioni precedenti, ha ora la struttura formale che la tesi richiedeva. Non è una celebrazione indiscriminata delle impossibilità: è una teoria precisa delle condizioni in cui le impossibilità strutturali generano strutture nuove, della forma in cui questa generazione avviene, e della misura — il surplus generativo  $\Gamma(I, S)$  — che ne quantifica l'intensità.

Il panorama che la teoria apre è questo. Il territorio del pensiero — lo spazio dei sistemi auto-rappresentazionali sufficientemente ricchi da incontrare le impossibilità del PAI — non è uno spazio liscio e uniforme. È un paesaggio con curvatura: le regioni di confine, dove le impossibilità strutturali definiscono le loro superfici, hanno massima curvatura informativa — sono i luoghi di massima densità di biforcazioni potenziali, dove il surplus generativo è più ricco e dove le condizioni del PCC possono più facilmente essere soddisfatte. Le regioni interne, lontane dai confini, sono le regioni di stabilità — dove il sistema opera comodamente dentro la propria struttura, senza incontrare il bordo. Le regioni esterne, oltre i confini, sono le regioni dei metalinguaggi — dove le impossibilità del livello precedente sono già state integrate e il sistema può trattarle come oggetti.

I sistemi più interessanti — le teorie più ricche, i cervelli più capaci, le scienze più produttive — sono i sistemi al confine o in prossimità dei confini. Sono i sistemi che incontrano le impossibilità strutturali, le riconoscono come tali, persistono attraverso l'incontro, producono le distinzioni che le impossibilità rendono necessarie, e hanno la struttura per integrare queste distinzioni in nuove strutture. Il cervello umano adulto maturo opera esattamente in questa regione: incontra il confine sintattico-semanticamente del divario esplicativo (impossibilità di Famiglia IV), il confine dimensionale del binding problem (Famiglia II), il confine computazionale delle euristiche invece dell'ottimizzazione esatta (Famiglia III), il confine diagonale dell'introspezione parziale (Famiglia I), il confine irreversibile della memoria consolidata e dell'identità nel tempo (Famiglia V). Non in modo astratto o filosofico: in modo funzionale, quotidiano, strutturalmente necessario per il suo operare.

Comprendere le impossibilità strutturali non è quindi soltanto un esercizio tecnico nella fondazione della matematica o nella biologia teorica. È la costruzione della mappa del territorio in cui il pensiero si muove — una mappa che mostra non soltanto i confini ma anche le direzioni in cui il pensiero si espande quando li incontra, la forma delle strutture che questi confini generano, e le condizioni sotto cui questa generazione avviene produttivamente invece che dissolversi nel fallimento o nel collasso. Il prossimo capitolo ha il compito di mostrare che questi confini non sono cinque oggetti separati ma le facce di un sistema unitario — un

sistema chiuso le cui componenti si implicano reciprocamente e la cui struttura algebrica riflette la coerenza interna della teoria che il ciclo ha costruito.

---

# Capitolo XII

## Il sistema chiuso

La struttura delle implicazioni reciproche tra le cinque famiglie, la legge di conservazione della carica impossibilitale, e la topografia del territorio cognitivo

### Mappa epistemica — Capitolo XII

| Status     | Cod. | Claim                                                                                                                                                           |
|------------|------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO | D    | $F_1 \rightarrow F_2$ : via il teorema di invarianza del dominio di Brouwer (1912).                                                                             |
| DIMOSTRATO | D    | $F_4 \rightarrow F_5$ : via il principio di Landauer (verificato sperimentalmente).                                                                             |
| CONGETTURA | C    | $F_2 \rightarrow F_3$ : dipende da $P \neq NP$ (Tillmann-Pfetsch 2014 sulla NP-hardness di RIP).                                                                |
| CONGETTURA | C    | $F_3 \rightarrow F_4$ : dipende da $P \neq NP$ (teorema di Fagin 1974).                                                                                         |
| CONGETTURA | C    | $F_5 \rightarrow F_1$ : dipende dalla condizione di auto-rappresentazione. La struttura $\mathbb{Z}_5$ è una congettura strutturale, non un teorema dimostrato. |

### I. Il problema dell'unità e il metodo della sua soluzione

**[CONGETTURA FONDATA]**

Avvertenza epistemica sulla struttura  $\mathbb{Z}_5$ . Le cinque implicazioni che formano il ciclo hanno forza epistemica disomogenea:

- $F_1 \rightarrow F_2$ : dimostrata via teorema di Brouwer (1912) [D]
- $F_4 \rightarrow F_5$ : fisicamente fondata via Landauer, verificata sperimentalmente [D]
- $F_2 \rightarrow F_3$ : dipende da  $P \neq NP$  (Tillmann-Pfetsch 2014) [C]
- $F_3 \rightarrow F_4$ : dipende da  $P \neq NP$  (Fagin 1974) [C]
- $F_5 \rightarrow F_1$ : dipende dalla condizione di auto-rappresentazione sufficiente [C]

La struttura  $\mathbb{Z}_5$  è una congettura strutturale fondata, non un teorema matematico. Il lettore è invitato a trattarla come ipotesi di ricerca strutturalmente motivata.

Undici capitoli hanno costruito una tassonomia delle impossibilità strutturali. Il catalogo è esteso: ha attraversato la logica matematica, la biologia evolutiva, la neurobiologia cognitiva, la teoria della complessità computazionale, la filosofia della scienza, i sistemi di intelligenza artificiale. Le cinque famiglie — impossibilità diagonali, dimensionali, di complessità, di chiusura semantica, di irreversibilità —

sono state identificate, distinte, e analizzate ciascuna nella propria struttura specifica. i capitoli precedenti del ciclo ne hanno mostrato le manifestazioni in ciascuno dei domini esaminati. Il capitolo XI ha mostrato come ciascuna famiglia abbia una forma di costruttività condizionata.

Un catalogo, però, non è una teoria. Una teoria richiede non soltanto la classificazione degli oggetti in famiglie ma la comprensione delle relazioni tra le famiglie — la struttura che le connette non come aggregato accidentale ma come sistema. La tesi di questo capitolo conclusivo è che le cinque famiglie formano un sistema chiuso nel senso formale preciso: ogni famiglia implica le altre, in condizioni appropriate, e queste implicazioni formano una struttura algebrica coerente isomorfa al gruppo ciclico  $\mathbb{Z}_5$ . La chiusura del sistema non è un'aggiunta estetica alla tassonomia: è il risultato teorico centrale che trasforma il catalogo in teoria.

Prima di procedere, occorre però una chiarificazione metodologica indispensabile, che concerne il senso in cui si afferma che le famiglie si implicano reciprocamente. Non si afferma una versione forte: qualsiasi sistema  $S$  che incontra la famiglia  $F_i$  incontra necessariamente tutte le altre famiglie, in qualsiasi condizione, senza eccezioni. Questa versione forte è falsa per controesempio: un sistema molto semplice che incontra un'impossibilità dimensionale (perché tenta di proiettare  $\mathbb{R}^3$  in  $\mathbb{R}^1$ ) potrebbe non avere abbastanza struttura auto-referenziale da incontrare un'impossibilità diagonale, se è troppo semplice per contenere la codifica gödeliana. Un sistema che incontra il ratchet di Müller a causa di dimensione della popolazione e tasso di mutazione sfavorevoli potrebbe non incontrare il frame problem se il suo dominio cognitivo è abbastanza ristretto da evitare la  $\Pi_2^P$ -completezza della Circumscription.

La versione difendibile — la versione per cui si offre dimostrazione — è una versione condizionale: per sistemi sufficientemente ricchi nel senso del PAI (sistemi che contengono l'equivalente dell'aritmetica del primo ordine e possono fare auto-riferimento nella misura necessaria), incontrare una delle cinque famiglie comporta, nelle condizioni strutturali appropriate, incontrare le famiglie lungo il ciclo. Questa versione condizionale è la versione rilevante: i sistemi cognitivi, biologici, e computazionali che questo ciclo ha analizzato — cervelli, sistemi evolutivi, sistemi di intelligenza artificiale, istituzioni scientifiche — sono tutti

sistemi sufficientemente ricchi. Per questi sistemi, la versione condizionale è operativamente equivalente alla versione forte.

Il metodo della dimostrazione non è assiomatico-deduttivo nel senso di partire da assiomi e ricavare teoremi per passaggi formali. È un metodo ibrido che usa strumenti matematici precisi (il teorema di invarianza del dominio di Brouwer, la NP-hardness della verifica RIP, il teorema di Fagin, il principio di Landauer con la sua verifica sperimentale, la costruzione diagonale per l'irreversibilità), applica questi strumenti a strutture formalmente specificate, e ricava implicazioni condizionali con il grado di rigore che i singoli passaggi permettono. Dove il pascapitolo è formalmente preciso, la dimostrazione è completa. Dove il pascapitolo usa risultati dipendenti da congetture non dimostrate (tipicamente  $P \neq NP$ ), questo è esplicitamente dichiarato. Dove il pascapitolo è argomentativo invece che formale, è presentato come tale. La trasparenza epistemologica sul grado di rigore di ciascun pascapitolo è parte integrante del metodo.

## **II. L'implicazione $F_1 \rightarrow F_2$ : come l'impossibilità diagonale genera l'impossibilità dimensionale**

Il primo anello del ciclo connette la Famiglia I (impossibilità diagonali) alla Famiglia II (impossibilità dimensionali). L'implicazione afferma che un sistema auto-rappresentazionale che incontra un'impossibilità diagonale incontra anche, nelle condizioni appropriate, un'impossibilità dimensionale. Il meccanismo è il seguente: l'impossibilità diagonale afferma che la funzione di auto-rappresentazione  $\varphi: \Sigma \rightarrow M^\Sigma$  non è suriettiva; la non-suriettività, in presenza di struttura metrica compatibile, implica una discrepanza dimensionale tra lo spazio sintattico  $\Sigma$  e lo spazio semantico  $M$ .

La formulazione rigorosa usa il teorema di invarianza del dominio, dimostrato da Brouwer nel 1912. Il teorema afferma: se  $f: U \rightarrow \mathbb{R}^n$  è una mappa continua e iniettiva definita su un aperto  $U \subseteq \mathbb{R}^n$ , allora  $f(U)$  è aperto in  $\mathbb{R}^n$ . Il contrappositivo è: se  $f: U \rightarrow V$  è una mappa continua con  $U$  aperto in  $\mathbb{R}^n$  e  $V$  in  $\mathbb{R}^m$ , e se  $m < n$ , allora  $f$  non può essere iniettiva — deve collassare almeno alcune coppie di punti.

Tradotto nel linguaggio del sistema chiuso: sia  $\Sigma$  lo spazio sintattico del sistema  $S$ , con struttura topologica indotta dalla struttura dei simboli e delle regole di

formazione, e sia  $M^\Sigma$  lo spazio semantico delle funzioni da  $\Sigma$  a  $M$ . Se  $\varphi: \Sigma \rightarrow M^\Sigma$  non è suriettiva (impossibilità diagonale) ma è iniettiva (nel senso che rappresentazioni sintattiche distinte producono funzioni semantiche distinte), allora per il teorema di invarianza del dominio  $\varphi(\Sigma)$  non può essere aperto in  $M^\Sigma$  se la dimensione di  $\Sigma$  è uguale a quella di  $M^\Sigma$  — il che, combinato con il fatto che  $\varphi(\Sigma)$  è compatto (se  $\Sigma$  è compatto), implica che  $\varphi(\Sigma)$  è un sottoinsieme proprio e chiuso di  $M^\Sigma$ , ossia che la dimensione di  $\Sigma$  è strettamente inferiore a quella di  $M^\Sigma$ .

Formalmente: siano  $\Sigma$  e  $M^\Sigma$  dotati di strutture metriche compatibili con la struttura logica del sistema (la metrica di Hamming sui codici gödeliani, o la norma euclidea sugli spazi vettoriali di rappresentazione). Se  $\dim(\Sigma) = \dim(M^\Sigma)$  e  $\varphi$  è continua e iniettiva, il teorema di invarianza del dominio implica che  $\varphi$  è omeomorfismo aperto, e quindi suriettiva su  $M^\Sigma$  connesso — contraddizione con l'impossibilità diagonale. Quindi:

Impossibilità diagonale + struttura metrica compatibile + iniettività di  $\varphi \rightarrow \dim(\Sigma) < \dim(M^\Sigma)$

Il lato destro è la struttura dell'impossibilità dimensionale: la dimensione dello spazio di rappresentazione sintattica è strettamente inferiore alla dimensione dello spazio semantico. Non si possono rappresentare fedelmente tutti gli oggetti semantici in uno spazio sintattico di dimensione insufficiente — l'impossibilità dimensionale nel senso della Famiglia II.

L'istanza biologica di questa implicazione connette il paradosso di Rogers (impossibilità della Famiglia I nel dominio culturale) al binding problem (impossibilità della Famiglia II nel dominio cognitivo). Il paradosso di Rogers afferma che la funzione di trasmissione culturale  $\varphi: \text{Comportamento} \rightarrow \text{Valore\_adattativo}$  non è suriettiva — esistono valori adattativi (adattamenti all'ambiente specifico) che non sono accessibili attraverso la sola copia del comportamento osservabile. Questa è un'impossibilità diagonale nella sua struttura formale: lo spazio dei comportamenti osservabili (la sintassi culturale) non copre lo spazio dei valori adattativi (la semantica culturale). Per l'implicazione dimostrata, questo implica che la dimensione dello spazio dei comportamenti è inferiore alla dimensione dello spazio dei valori — esattamente la struttura del binding problem nel dominio della cognizione culturale: non si possono codificare tutti i valori

adattativi nella struttura finita dei comportamenti osservabili e trasmissibili, e la discrepanza dimensionale produce la perdita di informazione che il paradosso di Rogers descrive.

### III. L'implicazione $F_2 \rightarrow F_3$ : come l'impossibilità dimensionale genera l'impossibilità di complessità

Il secondo anello connette la Famiglia II alla Famiglia III. L'implicazione afferma che un sistema che incontra un'impossibilità dimensionale — che deve operare su un dominio di dimensione superiore a quella della propria struttura di rappresentazione — incontra necessariamente un'impossibilità di complessità: il problema di trovare la rappresentazione ottimale è computazionalmente intrattabile.

La dimostrazione usa il risultato di Tillmann e Pfetsch (2014) sulla NP-hardness della verifica della Restricted Isometry Property (RIP). La RIP è la proprietà che caratterizza le proiezioni quasi-isometriche nel senso di Johnson-Lindenstrauss: una matrice  $A \in \mathbb{R}^{n \times m}$  soddisfa  $\text{RIP}(k, \delta)$  se  $(1-\delta)\|x\|^2 \leq \|Ax\|^2 \leq (1+\delta)\|x\|^2$  per ogni vettore  $k$ -sparso  $x$ . Trovare una matrice  $A$  che soddisfa  $\text{RIP}(k, \delta)$  con parametri ottimali (il minimo  $n$  dato  $m, k, \delta$ ) è il problema dell'ottimizzazione della proiezione dimensionale.

Tillmann e Pfetsch dimostrano che verificare se una data matrice binaria soddisfa  $\text{RIP}(k, \delta)$  è NP-hard per  $k = \Omega(\log m)$ . Il problema di ottimizzazione — trovare la matrice ottimale invece di soltanto verificarla — è quindi almeno altrettanto difficile, e in genere più difficile (la ricerca nello spazio delle matrici con la garanzia di ottimalità è computazionalmente più costosa della sola verifica). Formalmente: sia  $\text{OPT\_RIP}$  il problema "dato  $m, n, k, \delta$ , trovare  $A \in \mathbb{R}^{n \times m}$  che soddisfi  $\text{RIP}(k, \delta)$  e minimizzi  $n$ ".  $\text{OPT\_RIP}$  è NP-hard per  $k = \Omega(\log m)$ , e quindi appartiene alla Famiglia III dell'impossibilità di complessità.

Il condizionale che qualifica l'implicazione è  $P \neq \text{NP}$ : se  $P = \text{NP}$ , il problema  $\text{OPT\_RIP}$  sarebbe risolvibile in tempo polinomiale, e l'implicazione  $F_2 \rightarrow F_3$  verrebbe meno. Come già discusso per la Famiglia III, si tratta di una congettura fortemente sostenuta ma non dimostrata, e il carattere condizionale dell'implicazione deve essere esplicitamente riconosciuto.

L'istanza biologica è il frame problem come conseguenza del binding problem. Il sistema cognitivo che deve operare in un mondo con  $m$  feature usando  $n < m$  dimensioni rappresentative (binding problem, Famiglia II) deve affrontare il problema di aggiornare la propria rappresentazione del mondo dopo ogni azione — decidere quali delle  $m$  feature cambiano e quali restano invariate. Questo è il frame problem, e la sua  $\Pi_2^P$ -completezza (dimostrata da Liberatore e Schaerf 1998) è esattamente un'impossibilità di Famiglia III. La connessione non è una coincidenza: la  $\Pi_2^P$ -completezza emerge perché il sistema deve valutare condizioni  $\forall \exists$  (per ogni modello minimale, esiste la conseguenza) su uno spazio di modelli la cui complessità è determinata dal numero di feature  $m$  — esattamente il parametro che la discrepanza dimensionale  $n < m$  ha reso problematico. L'impossibilità dimensionale ( $n < m$ ) è la causa strutturale dell'impossibilità computazionale (la complessità della Circumscription scala con  $m$ ).

#### **IV. L'implicazione $F_3 \rightarrow F_4$ : come l'impossibilità di complessità genera l'impossibilità di chiusura semantica**

Il terzo anello connette la Famiglia III alla Famiglia IV. L'implicazione afferma che un sistema computazionalmente limitato che incontra un'impossibilità di complessità — che deve approssimare invece di ottimizzare esattamente — incontra necessariamente un'impossibilità di chiusura semantica: non può verificare dall'interno se le proprie rappresentazioni del mondo sono vere.

La dimostrazione usa il teorema di Fagin (1974) e la sua estensione da Stockmeyer (1976). Il teorema di Fagin afferma  $NP = \exists SO$ : un problema appartiene a NP se e soltanto se è definibile da una sentenza del secondo ordine con un quantificatore esistenziale sul simbolo di relazione. Estendendo, i livelli superiori della gerarchia polinomiale corrispondono a livelli della gerarchia delle formule del secondo ordine:  $\Sigma^k P$  corrisponde alle formule con  $k$  quantificatori alternanti del secondo ordine. In particolare,  $\Pi_2^P$  (il livello della Circumscription e del frame problem) corrisponde alle formule  $\forall \exists SO$  — formule che quantificano universalmente su relazioni del secondo ordine e poi esistenzialmente.

La semantica di un problema  $\Pi_2^P$  richiede, per essere completamente caratterizzata, la valutazione di proprietà del secondo ordine — proprietà che

appartengono a un livello della gerarchia tarskiana superiore al livello sintattico del sistema. Un sistema  $S$  limitato a computazioni in tempo polinomiale (il livello  $P$  — puramente sintattico nel senso delle operazioni sui simboli) non può esprimere completamente la semantica di un problema  $\Pi_2^P$  dall'interno della propria struttura: la relazione di soddisfazione per le formule  $\forall \exists$ SO richiede accesso alla struttura quantificata del secondo ordine, che è per definizione al di là del livello sintattico di  $S$ .

Questa è esattamente l'impossibilità di chiusura semantica della Famiglia IV: il sistema non può definire il predicato di verità per il proprio dominio all'interno della propria struttura sintattica, non per le ragioni del teorema di Tarski (che opera a prescindere dai vincoli computazionali) ma per ragioni computazionali — la valutazione del predicato di verità richiederebbe risorse che eccedono quelle disponibili al sistema. Il gap di complessità tra  $P$  e  $\Pi_2^P$  è un gap semantico nel senso del capitolo III di questo ciclo: lo spazio sintattico ( $P$ ) è di dimensione strettamente inferiore allo spazio semantico ( $\Pi_2^P$ ) nella gerarchia delle formule del secondo ordine.

Il condizionale rimane  $P \neq NP$ , e si applica anche qui.

Il cerchio biologico si chiude in modo particolarmente istruttivo per questa implicazione. Il choice blindness (Johansson et al. 2005) è il fenomeno per cui i soggetti, interrogati sulle proprie scelte dopo averle effettuate, forniscono spiegazioni che non corrispondono ai processi reali che le hanno generate — anzi, non rilevano nemmeno quando le scelte che vengono presentate loro come proprie sono state sostituite con scelte diverse. Interpretato nel linguaggio delle famiglie, il choice blindness è la manifestazione dell'impossibilità di chiusura semantica (Famiglia IV) generata dall'impossibilità computazionale (Famiglia III): il sistema cognitivo usa il Sistema 1 (euristiche veloci, approssimate, sub-simboliche) per effettuare le scelte — un sistema che opera a livello computazionale non accessibile al Sistema 2 (ragionamento esplicito, verbale, più lento). Il Sistema 2, quando deve spiegare la scelta del Sistema 1, non ha accesso ai processi reali (impossibilità di chiusura semantica: il predicato di verità per i processi del Sistema 1 non è definibile all'interno del Sistema 2) e costruisce quindi narrazioni post-hoc che possono essere sistematicamente errate. L'impossibilità computazionale del Sistema 1 (che usa euristiche invece dell'ottimizzazione esatta per ragioni di

Famiglia III) genera l'opacità semantica del Sistema 1 rispetto al Sistema 2 (Famiglia IV).

## **V. L'implicazione $F_4 \rightarrow F_5$ : come l'impossibilità di chiusura semantica genera l'impossibilità di irreversibilità**

Il quarto anello connette la Famiglia IV alla Famiglia V. L'implicazione afferma che un sistema fisicamente realizzato che incontra un'impossibilità di chiusura semantica — che ha un gap tra la struttura sintattica e la struttura semantica del proprio dominio — e che tenta di ridurre questo gap attraverso un processo di apprendimento o sviluppo, incontra necessariamente irreversibilità.

La dimostrazione usa il principio di Landauer (1961) nella sua formulazione termodinamica. Il principio afferma che qualsiasi operazione computazionale logicamente irreversibile — qualsiasi cancellazione di 1 bit di informazione — dissipa almeno  $k_B T \ln 2$  di energia nell'ambiente, dove  $k_B$  è la costante di Boltzmann e  $T$  la temperatura assoluta. Questo non è un risultato puramente matematico: è un teorema di fisica termodinamica, verificato sperimentalmente da Bérut et al. (2012) con una precisione del 4% rispetto al valore teorico.

Sia  $S$  un sistema con gap semantico  $\text{Gap}(\Sigma_S, M_S) = H(M_S) - I(M_S; \Sigma_S) > 0$ , fisicamente realizzato su un substrato materiale a temperatura  $T > 0$ . Un processo di apprendimento  $T: S \rightarrow S'$  che riduce il gap deve aumentare  $I(M_S; \Sigma_S)$  — deve portare più informazione su  $M_S$  nella struttura sintattica  $\Sigma_S$ . Ci sono due modi in cui questo può avvenire:

Il primo modo è l'espansione di  $\Sigma_S$ : si aggiunge struttura sintattica nuova (nuove sinapsi, nuovi parametri, nuovi assiomi) che cattura informazione semantica non catturata prima. L'espansione richiede, fisicamente, un costo metabolico o computazionale. La struttura aggiunta non è recuperabile senza un ulteriore investimento di risorse — è incorporata nel substrato fisico in modo che rimuoverla richiederebbe lo stesso tipo di costo che costruirla. Questa è una forma di irreversibilità: il sistema non può "tornare" alla struttura precedente senza un costo paragonabile al costo di costruzione della struttura nuova.

Il secondo modo è la sostituzione: si rimuove struttura sintattica irrilevante (che non porta informazione su  $M_S$ ) e la si sostituisce con struttura sintattica

rilevante. La rimozione è la cancellazione di informazione nel senso del principio di Landauer: l'informazione sulla struttura rimossa viene cancellata, con dissipazione di almeno  $k_B T \ln 2$  per bit rimosso. Questa cancellazione è irreversibile: l'informazione dissipata sotto forma di calore nell'ambiente non è recuperabile (il secondo principio della termodinamica). Due stati di  $\Sigma_S$  che differivano soltanto per le strutture sintattiche rimosse producono lo stesso stato dopo la sostituzione — il processo  $T$  è non-iniettivo.

In entrambi i casi,  $T$  produce irreversibilità nel senso formale della Famiglia V. L'impossibilità di chiusura semantica ( $\text{Gap} > 0$ ) combinata con il tentativo di ridurre il gap su un substrato fisico a temperatura positiva implica necessariamente l'irreversibilità dei processi di apprendimento o sviluppo.

Questa implicazione è, tra le cinque, quella con il fondamento fisico più diretto. Il principio di Landauer non è una congettura matematica: è un risultato di fisica con verifica sperimentale. L'implicazione  $F_4 \rightarrow F_5$  non dipende da  $P \neq NP$  né da altre congetture non dimostrate: dipende dalla termodinamica, e la termodinamica è una delle teorie fisiche più solidamente verificate.

L'istanza biologica è la potatura sinaptica durante lo sviluppo come conseguenza dell'impossibilità di chiusura semantica nell'apprendimento. Il sistema nervoso in sviluppo ha un gap semantico — la struttura sinaptica non cattura completamente la semantica dell'ambiente in cui l'organismo deve operare. Il processo di sviluppo tenta di ridurre questo gap attraverso la potatura: le sinapsi debolmente attivate (che non portano informazione rilevante sull'ambiente) vengono eliminate, e le sinapsi fortemente attivate (che codificano le strutture ambientali rilevanti) vengono rafforzate. La potatura è l'eliminazione di struttura sintattica irrilevante — una cancellazione di informazione nel senso di Landauer. La dissipazione energetica è reale: ogni sinapsi eliminata richiede l'attivazione dei processi molecolari di degradazione (caspasi, autofagia sinaptica) che consumano ATP e disperdono calore. L'irreversibilità è reale: le sinapsi potate non vengono ristabilite senza attivazione specifica (come avviene in certi contesti patologici o farmacologici, ma non nel corso normale dello sviluppo). La potatura sinaptica è la manifestazione fisica, misurabile in joule, dell'implicazione  $F_4 \rightarrow F_5$ .

## VI. L'implicazione $F_5 \rightarrow F_1$ : come l'impossibilità di irreversibilità genera l'impossibilità diagonale

Il quinto anello chiude il ciclo, connettendo la Famiglia V alla Famiglia I. L'implicazione afferma che un sistema auto-rappresentazionale che ha subito processi irreversibili — la cui dinamica è non-iniettiva — incontra necessariamente un'impossibilità diagonale quando tenta di costruire una rappresentazione completa della propria storia.

Il meccanismo della dimostrazione usa la struttura della non-iniettività della dinamica e la costruzione diagonale di Cantor-Gödel applicata allo spazio delle traiettorie storiche.

Sia  $S$  un sistema con dinamica  $T: S \rightarrow S$  tale che  $T(s_1) = T(s_2) = s^*$  per qualche  $s_1 \neq s_2$  (non-iniettività, l'irreversibilità formale della Famiglia V). Sia  $s_{\text{curr}} = T^n(s_0)$  lo stato corrente del sistema dopo  $n$  passi di evoluzione. Il sistema tenta di costruire una funzione  $\varphi: S_{\text{curr}} \rightarrow \text{Traiettorie}^{\{S_{\text{curr}}\}}$  che rappresenti, a partire dallo stato corrente, tutte le possibili traiettorie storiche che hanno portato al sistema a  $s_{\text{curr}}$ .

Per la condizione di auto-rappresentazione — che il sistema sia sufficientemente ricco da codificare le proprie traiettorie come oggetti del dominio — si applica il lemma della diagonalizzazione: per qualsiasi funzione  $\varphi$  della forma data, esiste una traiettoria diagonale  $d$  che "differisce" da  $\varphi_{\{s_{\text{curr}}\}}$  nella dimensione dell'informazione cancellata dalla non-iniettività. La costruzione è la seguente: la non-iniettività implica che nel passato del sistema c'è almeno una biforcazione — un punto in cui due stati distinti si sono fusi in uno. L'informazione che distingueva quei due stati è stata cancellata nella transizione. La traiettoria diagonale  $d$  è la traiettoria che, ad ogni biforcazione del passato, prende il ramo opposto a quello indicato dall'assegnazione di  $\varphi$ . Poiché l'informazione sul ramo effettivamente percorso è stata cancellata,  $d$  non appartiene all'immagine di  $\varphi$ : qualsiasi assegnazione  $\varphi$  fa su  $s_{\text{curr}}$  non può includere  $d$  come una traiettoria possibile, perché per includerla dovrebbe avere l'informazione cancellata che distingue  $d$  dalla traiettoria effettivamente percorsa.

La conclusione è che  $\varphi$  non è suriettiva rispetto allo spazio delle traiettorie storiche effettivamente possibili — un'impossibilità diagonale nel senso della Famiglia I.

Il condizionale che qualifica questa implicazione è la condizione di auto-rappresentazione: il sistema deve essere sufficientemente ricco da codificare le proprie traiettorie come oggetti del dominio e da applicare il lemma della diagonalizzazione. Per i sistemi semplici che mancano di questa struttura (sistemi fisici non auto-rappresentazionali, come un cristallo o un gas), l'implicazione non si attiva. Per i sistemi cognitivi, biologici complessi, e computazionali ricchi, la condizione è soddisfatta.

L'istanza biologica è la ricostruzione filogenetica come caso di impossibilità diagonale generata dall'irreversibilità evolutiva. Il genoma attuale di una specie ( $s_{curr}$ ) è il risultato di innumerevoli biforcazioni evolutive — speciazioni, sostituzioni alleliche, eventi di deriva genetica — che hanno cancellato l'informazione sulle varianti non selezionate. Qualsiasi albero filogenetico ricostruito dal genoma attuale (una funzione  $\varphi$  che mappa il genoma corrente su storie evolutive possibili) è incompleto: esistono sempre alberi evolutivi alternativi (traiettorie diagonali  $d$ ) compatibili con i dati genomici correnti che non sono distinguibili sulla base del solo genoma attuale. I metodi filogenetici più sofisticati (massima verosimiglianza, inferenza bayesiana) producono distribuzioni di probabilità sugli alberi possibili — non alberi singoli — precisamente perché riconoscono implicitamente l'impossibilità diagonale della ricostruzione storica completa.

## VII. La struttura algebrica del sistema: $\mathbb{Z}_5$ e le sue proprietà

Le cinque implicazioni  $F_1 \rightarrow F_2 \rightarrow F_3 \rightarrow F_4 \rightarrow F_5 \rightarrow F_1$  formano un grafo orientato su cinque nodi. La struttura di questo grafo è isomorfa al grafo del generatore del gruppo ciclico  $\mathbb{Z}_5$  — il gruppo con cinque elementi  $\{0,1,2,3,4\}$  e addizione modulo 5, generato dall'elemento 1.

L'identificazione è formale: si mappano le famiglie su  $\mathbb{Z}_5$  attraverso  $F_k \leftrightarrow k-1$  (con  $F_1 \leftrightarrow 0$ ,  $F_2 \leftrightarrow 1$ , ...,  $F_5 \leftrightarrow 4$ ), e l'implicazione  $F_k \rightarrow F_{\{k+1 \bmod 5\}}$  corrisponde all'operazione del generatore  $\sigma = 1$  di  $\mathbb{Z}_5$ :  $\sigma(k) = k+1 \bmod 5$ . Il fatto che il ciclo di implicazioni si chiuda ( $F_5 \rightarrow F_1$  invece di generare una famiglia  $F_6$  distinta) corrisponde alla proprietà  $\sigma^5 = 0$  in  $\mathbb{Z}_5$ .

Questa struttura algebrica non è soltanto un'analogia elegante: ha conseguenze dimostrabili per il sistema delle impossibilità.

La semplicità di  $\mathbb{Z}_5$  è la prima conseguenza rilevante.  $\mathbb{Z}_5$  è un gruppo semplice — non ha sottogruppi propri non banali, perché 5 è primo. Questo significa che non esiste un sottoinsieme proprio non banale delle cinque famiglie che sia chiuso rispetto alle implicazioni senza coinvolgere le famiglie rimanenti. Non si possono isolare due, tre, o quattro famiglie in un sotto-sistema chiuso: qualsiasi coppia di famiglie che generi un ciclo di implicazioni finisce per coinvolgere tutte le rimanenti. Il sistema è minimalmente chiuso nel senso che 5 è il minimo numero di famiglie per cui la struttura ciclica forma un gruppo con questa proprietà di indivisibilità.

La ciclicità genera le implicazioni composte come conseguenza del teorema di Cayley applicato al gruppo ciclico. Poiché  $\mathbb{Z}_5$  è generato da  $\sigma = 1$ , ogni elemento è una potenza del generatore:  $2 = \sigma^2$ ,  $3 = \sigma^3$ ,  $4 = \sigma^4$ . Le implicazioni composte corrispondono alle potenze del generatore:

$F_1 \rightarrow_2 F_3$  (via  $F_2$ ): la composizione di  $F_1 \rightarrow F_2$  e  $F_2 \rightarrow F_3$  dà la transizione  $\sigma^2$  nel gruppo, cioè il salto di due posizioni nella catena. Concretamente: un sistema che incontra un'impossibilità diagonale incontra, attraverso la mediazione dell'impossibilità dimensionale, un'impossibilità di complessità.

$F_1 \rightarrow_3 F_4$  (via  $F_2, F_3$ ): il salto di tre posizioni, corrispondente a  $\sigma^3$ . Un sistema che incontra un'impossibilità diagonale incontra, attraverso le mediazioni delle Famiglie II e III, un'impossibilità di chiusura semantica.

$F_1 \rightarrow_4 F_5$  (via  $F_2, F_3, F_4$ ): il salto di quattro posizioni,  $\sigma^4$ . La catena completa porta dalla Famiglia I alla Famiglia V attraverso le tre famiglie intermedie.

La transitività delle implicazioni — il fatto che i salti composti siano anch'essi implicazioni valide per i sistemi appropriati — implica la totalità della chiusura: per qualsiasi sistema  $S$  sufficientemente ricco che incontra qualsiasi famiglia  $F_i$ ,  $S$  incontra (nelle condizioni strutturali appropriate) tutte e cinque le famiglie. Il sistema è totalmente connesso nella struttura di implicazioni.

Questa proprietà ha un'interpretazione concreta nei sistemi biologici complessi. Il cervello umano adulto incontra simultaneamente tutte e cinque le famiglie: l'impossibilità diagonale dell'introspezione parziale (non si può osservare completamente i propri processi cognitivi perché l'osservazione modifica i processi osservati), l'impossibilità dimensionale del binding problem (il numero di

feature nel mondo supera la dimensione dello spazio di rappresentazione neurale), l'impossibilità di complessità delle euristiche (le scelte ottimali richiedono computazioni NP-hard, quindi si usano approssimazioni), l'impossibilità di chiusura semantica del divario esplicativo (le strutture neurali non implicano la propria semantica fenomenica), e l'impossibilità di irreversibilità della memoria consolidata e dell'identità (le strutture apprese e cristallizzate durante lo sviluppo e l'esperienza non sono liberamente modificabili). Il fatto che il cervello incontri tutte cinque simultaneamente non è una coincidenza accidentale: è la conseguenza della struttura del sistema chiuso, che per un sistema sufficientemente ricco implica la totalità delle famiglie.

### **VIII. La legge di conservazione della carica impossibilitale: la parte dimostrabile**

il capitolo VII di questo ciclo ha proposto la congettura che le impossibilità si conservino — che ridurre l'intensità di una famiglia implichi necessariamente l'aumento dell'intensità di almeno un'altra. Questa congettura, nella sua forma generale, è ancora un programma di ricerca. Ma una versione parziale — la compensazione tra la Famiglia I e la Famiglia II — è dimostrabile con gli strumenti sviluppati nel ciclo.

Si definiscono le misure di intensità per le due famiglie nel contesto specifico di un sistema di rappresentazione con  $n$  dimensioni sintattiche e  $m$  oggetti semantici da rappresentare.

$I_2(S) = \max(0, \log_2 m - n)$  — la discrepanza informazionale tra il numero di oggetti distinti nel dominio semantico ( $\log_2 m$  bit sono necessari per identificare uno tra  $m$  oggetti) e la dimensione dello spazio sintattico ( $n$  dimensioni). Questa misura vale zero quando  $n \geq \log_2 m$  (il sistema ha abbastanza dimensioni per distinguere tutti gli oggetti) e cresce linearmente quando  $n < \log_2 m$ .

$I_1(S)$  — la misura dell'impossibilità diagonale — è più delicata da definire in modo compatibile con  $I_2$ . Nel contesto dei sistemi di rappresentazione con  $n$  dimensioni, la misura naturale è il logaritmo del numero di oggetti semantici che il sistema può formulare ma non "raggiungere" con la propria funzione di auto-rappresentazione. Per un sistema con  $n$  dimensioni sintattiche, il numero massimo di oggetti

distinguibili è  $2^n$  (il numero di configurazioni binarie di  $n$  dimensioni). Se il dominio semantico ha  $m$  oggetti, il numero di oggetti irraggiungibili è  $\max(0, m - 2^n)$ . Quindi:

$$I_1(S) = \max(0, \log_2 m - n) = I_2(S)$$

Nella definizione più naturale, le due misure coincidono — un risultato che potrebbe sembrare banale ma che ha la struttura giusta per enunciare la legge di conservazione.

Si considera una trasformazione  $T: S \rightarrow S'$  che aumenta la dimensione sintattica di  $\Delta$ :  $n' = n + \Delta$ . Allora:

$$I_2(S') = \max(0, \log_2 m - (n + \Delta)) = \max(0, I_2(S) - \Delta)$$

La dimensione aggiuntiva riduce l'impossibilità dimensionale di  $\Delta$ . Ma con la dimensione aggiuntiva, il sistema può ora formulare strutture auto-referenziali più ricche: la potenza espressiva del sistema aumenta, e quindi il numero di oggetti semantici che il sistema può "vedere" ma non raggiungere — il numero di proposizioni indimostrabile nella nuova teoria più forte — non decresce. Formalmente, per il teorema di incompletezza gödeliana applicato al sistema esteso: la teoria  $S'$  è più forte di  $S$  (ha più struttura auto-referenziale disponibile), e per ogni aumento della forza dimostrativa c'è un corrispondente aumento nel numero di proposizioni indecidibile. In termini di cardinalità, il numero di proposizioni vere ma indimostrabile in  $S'$  è almeno tanto quanto in  $S$  — e tipicamente, per le estensioni interessanti, è strettamente maggiore.

Questo produce la relazione di compensazione:

Riduzione di  $I_2$  di  $\Delta \leftrightarrow$  Non-riduzione di  $I_1$  (e tipicamente aumento)

La somma  $I_1 + I_2$  è non-decrescente rispetto alle trasformazioni che aumentano la dimensione sintattica. Questo è il teorema di conservazione parziale: la carica totale  $I_1 + I_2$  non si riduce — ridurre l'impossibilità dimensionale attraverso l'espansione sintattica non riduce la carica impossibilitante complessiva.

Una seconda istanza della legge di conservazione, anch'essa dimostrabile, riguarda le Famiglie IV e V. Il principio di Landauer fornisce la relazione quantitativa: ogni riduzione di  $\text{Gap}(\Sigma_S, M_S)$  di 1 bit (riduzione di  $I_4$ ) attraverso la cancellazione di struttura sintattica irrilevante produce una dissipazione di  $k_B T \ln 2$  di entropia

nell'ambiente (aumento di  $I_5$  misurato in unità di entropia termodinamica). La relazione è:

$$\Delta I_5 \geq -\Delta I_4 \cdot k_B \ln 2 \cdot T$$

che nella forma adimensionale (misurando entrambe in bit) diventa:

$$\Delta I_5 \geq -\Delta I_4$$

— ogni riduzione dell'impossibilità di chiusura semantica è compensata da almeno uguale aumento dell'impossibilità di irreversibilità. Questa è la versione termodinamica della legge di conservazione, e ha il fondamento fisico più solido di tutte: si riduce direttamente al principio di Landauer verificato sperimentalmente.

La versione generale della legge — che la carica totale  $\Psi(S) = \sum_i w_i \cdot I_i(S)$  sia conservata rispetto alle trasformazioni che preservano la struttura auto-rappresentazionale di  $S$  — rimane una congettura. La difficoltà è che la nozione precisa di "trasformazioni che preservano la struttura auto-rappresentazionale" non è ancora formalizzata in modo da permettere una dimostrazione generale. Le due istanze dimostrate ( $I_1+I_2$  e  $I_4+I_5$ ) sono le evidenze più solide disponibili che la legge generale valga, e sono sufficienti a dare contenuto empirico alla congettura.

## IX. La mappa del possibile come inversione formale della mappa dell'impossibile

La tesi che percorre l'intero ciclo — il confine dell'impossibile è anche la mappa del possibile — può ora essere formulata con precisione formale, non come affermazione retorica.

Sia  $S$  lo spazio di tutti i sistemi auto-rappresentazionali, dotato della topologia indotta dalla distanza di Kolmogorov  $d_K(S_1, S_2) = \max(K(S_1|S_2), K(S_2|S_1))$ . Per ciascuna famiglia  $F_i$ , la superficie di confine  $C_{-}\{F_i\}$  è il sottoinsieme di  $S$  dei sistemi che sono esattamente al confine dell'impossibilità  $F_i$  — sistemi abbastanza ricchi da formulare l'impossibilità, non abbastanza da avere già integrato la struttura superiore.

La decomposizione dello spazio  $S$  rispetto a  $C_{-}\{F_i\}$  è:

$$S = S_{-}(F_i) \sqcup C_{-}\{F_i\} \sqcup S_{+}(F_i)$$

dove  $S_-(F_i)$  è l'insieme dei sistemi troppo semplici per incontrare  $F_i$ , e  $S_+(F_i)$  è l'insieme dei sistemi che hanno già integrato  $F_i$  al livello superiore.

Questa decomposizione non è simmetrica.  $S_-(F_i)$  è la regione di semplicità — sistemi che non possono formulare le domande a cui  $F_i$  risponde negativamente.  $S_+(F_i)$  è la regione dei metalinguaggi — sistemi che trattano  $F_i$  come un oggetto nel proprio dominio.  $C_{-}\{F_i\}$  è la regione di massima tensione cognitiva — sistemi che formulano le domande ma non le possono risolvere dall'interno.

La struttura di implicazioni del sistema chiuso implica che per i sistemi sufficientemente ricchi, le cinque superfici di confine sono simultaneamente attive. Il sistema di sistemi al confine di tutte le famiglie è:

$$C^* = \bigcap_{i=1}^5 C_{-}\{F_i\} \text{ (o, più realisticamente, } \bigcap_i U_{-\epsilon}(C_{-}\{F_i\}) \text{ per qualche } \epsilon > 0)$$

dove  $U_{-\epsilon}$  è il  $\epsilon$ -vicinato. I sistemi in  $C^*$  sono i sistemi che abitano simultaneamente tutti i cinque confini — che incontrano tutte le cinque impossibilità e che le incontrano nelle condizioni appropriate per la costruttività del PCC. Questi sono i sistemi con la massima ricchezza cognitiva accessibile: non soltanto perché incontrano il massimo numero di impossibilità, ma perché le impossibilità che incontrano sono le fonti di costruttività più ricche — ciascuna genera, nelle condizioni appropriate, strutture superiori.

La mappa del possibile è la mappa topologica di  $C^*$  e delle sue connessioni con i livelli superiori attraverso le ascese gödeliane. Non è una mappa delle regioni in cui si opera senza incontrare confini (che sarebbero  $S_-(F_i)$  — le regioni di semplicità): è la mappa del confine stesso e delle direzioni in cui i confini si aprono verso il territorio superiore.

La struttura dei livelli nel fibrato cognitivo è la seguente. Per ogni ordinale  $\alpha$ , sia  $Cat(\alpha)$  la categoria dei sistemi cognitivi con profondità tarskiana  $\alpha$  — sistemi al livello  $\alpha$  nella gerarchia dei metalinguaggi. I sistemi in  $Cat(\alpha)$  incontrano le loro impossibilità strutturali, che possono generare sistemi in  $Cat(\alpha+1)$  attraverso le ascese gödeliane appropriate. Il fibrato cognitivo è:

$$F = \bigcup_{\alpha \in ORD\_costruttivi} Cat(\alpha)$$

con struttura di connessione data dal funtore di ascesa gödeliana  $G: Cat(\alpha) \rightarrow Cat(\alpha+1)$ .

Ogni livello  $\text{Cat}(\alpha)$  ha la propria copia del sistema chiuso delle cinque famiglie — il sistema  $\mathbb{Z}_5$  delle implicazioni si ripete a ogni livello della gerarchia, con la stessa struttura locale ma con oggetti diversi (le impossibilità del livello  $\alpha$  sono diversi dalle impossibilità del livello  $\alpha+1$ , anche se hanno la stessa struttura formale). La topografia del territorio cognitivo è quindi quella di un fibrato stratificato: infiniti livelli sovrapposti, ciascuno con cinque superfici di confine, tutti connessi attraverso le ascese gödeliane.

La proprietà più importante di questa struttura è la sua apertura verso l'alto: il fibrato non ha un livello supremo. Per qualsiasi livello  $\alpha$ , il PAI si applica ai sistemi di  $\text{Cat}(\alpha)$ : esiste un sistema auto-referenziale in  $\text{Cat}(\alpha)$  con proprie impossibilità strutturali, che le ascese appropriate portano a  $\text{Cat}(\alpha+1)$ . Non si raggiunge mai un livello in cui tutte le impossibilità sono integrate e nessuna nuova impossibilità emerge. Il territorio cognitivo è aperto verso l'alto nel senso topologico preciso: per ogni sistema, esistono sistemi più ricchi che lo contengono come oggetto del proprio dominio.

## **X. Il PAI applicato al ciclo: la coerenza dell'auto-confutazione**

Un ciclo teorico sulle impossibilità strutturali che non incontrasse le proprie impossibilità strutturali sarebbe formalmente incoerente — affermerebbe l'universalità del PAI ed ne sarebbe inspiegabilmente esente. Questo capitolo si chiude quindi con l'applicazione del PAI al ciclo stesso, come atto non di modestia ma di coerenza.

Le cinque impossibilità del ciclo sono reali e identificabili con precisione.

L'impossibilità diagonale del ciclo è la seguente. La tesi che le cinque famiglie siano complete — che coprano tutte le impossibilità strutturali rilevanti per i sistemi cognitivi, biologici, e computazionali — non è dimostrabile all'interno del sistema concettuale del ciclo. Per il lemma della diagonalizzazione applicato alla funzione che mappa i sistemi concettuali sulle tassonomie di impossibilità che possono generare, esiste sempre una struttura di impossibilità che "differisce" da ciascuna delle cinque famiglie nelle dimensioni rilevanti. Il ciclo non può affermare la propria completezza tassonomica dall'interno: questa affermazione richiederebbe un metalinguaggio esterno al ciclo. Le cinque famiglie sono la proposta più

economica e strutturalmente coerente disponibile al momento della scrittura, ma non sono dimostrativamente complete.

L'impossibilità dimensionale del ciclo è che il testo è una proiezione di dimensione finita di un territorio concettuale di dimensione superiore. Numerosi fenomeni che appartengono alla struttura delle impossibilità strutturali — le impossibilità nella fisica quantistica (il problema della misurazione come forma di chiusura semantica, l'entanglement come impossibilità di separazione locale), le impossibilità nelle scienze sociali (il problema ermeneutico come forma del gap semantico, l'impossibilità della traduzione completa di Quine come diagonalizzazione nell'ambito linguistico), le impossibilità nella teoria della comunicazione (i limiti di capacità di Shannon come forma di impossibilità dimensionale per i canali fisici) — non sono stati sviluppati con la profondità che meriterebbero. Il ciclo ha proiettato il territorio in undici dimensioni tematiche: la fedeltà della proiezione è alta nelle dimensioni scelte, inevitabilmente approssimata in quelle non scelte.

L'impossibilità di complessità del ciclo è che la costruzione sistematica di tutte le relazioni tra le cinque famiglie — tutti i possibili sistemi S che le incontrano, tutte le possibili condizioni del PCC, tutte le possibili istanze delle implicazioni — è computazionalmente intrattabile. Il ciclo ha selezionato le istanze paradigmatiche e le implicazioni principali, ma non può pretendere di avere esplorato lo spazio completo delle relazioni. La selezione è stata guidata dalla struttura teorica, non dall'enumerazione esaustiva.

L'impossibilità di chiusura semantica del ciclo è che i concetti centrali — "impossibilità strutturale", "confine sintassi-semantica", "sistema auto-rappresentazionale" — hanno una semantica che eccede la loro definizione sintattica nei testi. Il significato di "impossibilità strutturale" in questo ciclo è determinato non soltanto dalle definizioni formali ma dall'intero contesto della ricerca matematica, biologica, e filosofica in cui il ciclo si inserisce — un contesto che non è catturabile nei testi stessi. Un lettore che conosce questo contesto capirà i concetti in un modo che un lettore che non lo conosce non raggiungerà soltanto attraverso le definizioni. Il gap semantico del ciclo rispetto ai suoi lettori è reale e irriducibile.

L'impossibilità di irreversibilità del ciclo è che esso è irreversibilmente situato nel proprio momento storico. I teoremi citati, le strutture biologiche analizzate, i

sistemi computazionali discussi riflettono lo stato delle discipline al momento della scrittura. Nuovi teoremi di impossibilità (nel senso di incompletezza di nuovi sistemi formali, di nuovi lower bound computazionali, di nuove strutture biologiche che mostrano forme di irreversibilità o di auto-riferimento non analizzate qui) cambieranno il paecapitolo senza che il ciclo possa incorporarli retroattivamente. Il ciclo è un documento storico della teoria delle impossibilità strutturali in un certo momento — non un documento eterno.

Il riconoscimento di queste impossibilità non è la sconfessione del ciclo: è la dimostrazione della propria coerenza. Un ciclo che non incontrasse le proprie impossibilità strutturali sarebbe o troppo semplice (e quindi non sufficientemente ricco da porre le domande che il PAI rende irrisolvibili) o incoerente con la propria tesi centrale. Il ciclo incontra le proprie impossibilità perché è sufficientemente ricco: è nella regione  $C^*$  del proprio spazio concettuale. Questo significa che il surplus generativo del ciclo — ciò che non è già contenuto nel ciclo ma che l'incontro con le proprie impossibilità rende possibile — è positivo. Il ciclo non conclude: indica le direzioni in cui il pensiero può continuare.

## **XI. La struttura conclusiva: il confine come condizione, non come fine**

Il ciclo ha costruito una topografia. Non una lista di ciò che non si può fare — una topografia: la descrizione della struttura dello spazio in cui il pensiero, la vita e la computazione operano, con le sue regioni di stabilità, i suoi confini di impossibilità, le sue direzioni di ascesa verso livelli superiori.

La struttura topografica ha cinque dimensioni principali, corrispondenti alle cinque famiglie. Nella dimensione dell'auto-riferimento diagonale (Famiglia I), i sistemi si stratificano secondo la propria profondità nel processo di auto-applicazione: i sistemi più semplici non possono applicarsi a se stessi, i sistemi intermedi incontrano le impossibilità gödeliane dell'auto-applicazione, i sistemi superiori le hanno integrate nei propri metalinguaggi. Nella dimensione della fedeltà rappresentativa (Famiglia II), i sistemi si stratificano secondo il rapporto tra la dimensione del proprio spazio di rappresentazione e la dimensione del proprio dominio semantico: i sistemi con rappresentazione esatta non hanno bisogno di approssimazione, i sistemi con discrepanza dimensionale sono nel regime di JL, i

sistemi che hanno risolto il problema della compressione ottimale abitano il confine. Nella dimensione della profondità computazionale (Famiglia III), i sistemi si stratificano secondo il livello della gerarchia polinomiale dei problemi che affrontano: sistemi in P, sistemi in NP, sistemi a livelli superiori. Nella dimensione della profondità semantica (Famiglia IV), i sistemi si stratificano secondo il proprio livello nella gerarchia tarskiana dei metalinguaggi. Nella dimensione della storia irreversibile (Famiglia V), i sistemi si stratificano secondo la quantità di processi irreversibili accumulati nel tempo.

Ogni sistema cognitivo occupa una posizione in questo spazio a cinque dimensioni. La mappa del possibile è la topografia di questo spazio — la descrizione di come le cinque dimensioni si connettono (attraverso il sistema di implicazioni  $\mathbb{Z}_5$ ), di come i confini si strutturano (attraverso le superfici  $C_{\{F_i\}}$ ), e di come il territorio si stratifica verso l'alto (attraverso il fibrato delle categorie cognitive).

I sistemi biologici più complessi — il cervello umano, gli ecosistemi maturi, le comunità scientifiche — sono sistemi in  $C^*$ : abitano simultaneamente tutti e cinque i confini. Non per caso, ma per necessità strutturale: sono abbastanza ricchi da incontrare le impossibilità di tutte le famiglie, e questa ricchezza è la condizione della loro capacità di generare novità. Il surplus generativo  $\Gamma(I, S)$  è massimo nei sistemi in  $C^*$  — sono i sistemi da cui le strutture nuove emergono, quando le condizioni del PCC sono soddisfatte.

La mappa non è una mappa delle limitazioni. È una mappa della struttura del territorio — del modo in cui il territorio è organizzato, delle direzioni in cui si espande, dei confini che definiscono la transizione tra livelli. Un esploratore che conosce la topografia del territorio che percorre non è più limitato dalla topografia: sa dove si trovano i valichi, dove le correnti, dove i picchi da cui si vede più lontano. Il confine dell'impossibile — mappato con la precisione che il ciclo ha tentato di costruire — è questo: il sistema di indicazioni che permette di navigare il territorio cognitivo con il massimo di efficacia, senza perdere tempo in direzioni precluse e senza sprecare risorse in tentativi di attraversare confini che non si attraversano. Il pensiero che conosce le proprie impossibilità strutturali non è un pensiero limitato. È un pensiero orientato.

Appendici tecniche

---

**Mappa epistemica — Capitolo XIII — Nuovi Domini: Fisica e Comunicazione**

| Status          | Cod. | Claim                                                                                                                           |
|-----------------|------|---------------------------------------------------------------------------------------------------------------------------------|
| DIMOSTRATO      | D    | Teorema di Shannon sulla capacità del canale: $C = B \log_2(1 + S/N)$ . Limite strutturale, non contingente.                    |
| DIMOSTRATO      | D    | Principio di indeterminazione di Heisenberg: $\Delta x \Delta p \geq \hbar/2$ .                                                 |
| DIMOSTRATO      | D    | Teorema di non-clonazione quantistica (Wootters-Zurek 1982).                                                                    |
| DIMOSTRATO      | D    | Teorema di Bell (1964; verifica loop-hole-free: Hensen et al. 2015).                                                            |
| CONGETTURA      | C    | Il problema della misurazione quantistica come impossibilità di chiusura semantica tra meccanica quantistica e classica.        |
| INTERPRETAZIONE | I    | L'entanglement come binding dimensionale negli spazi di Hilbert; la decoerenza come versione quantistica del ratchet di Müller. |

# Capitolo XIII

## Nuovi domini: le impossibilità nella fisica e nella comunicazione

### L'indeterminazione quantistica, il limite di Shannon, le impossibilità distribuite e la struttura delle impossibilità nella fisica fondamentale

#### I. Perché un nuovo capitolo

L'analisi condotta nei dodici capitoli precedenti si è concentrata su quattro grandi aree: logica matematica e teoria della computazione, biologia evolutiva e neurobiologia, filosofia della scienza, e intelligenza artificiale. Due domini di grande rilevanza per la teoria delle impossibilità strutturali sono stati trattati soltanto marginalmente: la fisica fondamentale — in particolare la meccanica quantistica — e la teoria della comunicazione. Questo capitolo colma questa lacuna, identificando le strutture di impossibilità in questi domini e mostrandone le connessioni con le cinque famiglie.

La fisica fondamentale è il dominio in cui le impossibilità strutturali si manifestano con il massimo di precisione formale: la fisica è la disciplina che ha sviluppato il più alto grado di formalizzazione matematica dei propri risultati. Le impossibilità fisiche — il principio di indeterminazione, il teorema di non-clonazione, il teorema di Bell — non sono artefatti di teorie incomplete ma risultati dimostrati con il medesimo rigore dei teoremi di Gödel e Turing. La loro inserzione nella tassonomia delle cinque famiglie non è un'estensione speculativa: è un'identificazione formale.

## II. La teoria di Shannon e i limiti della comunicazione (Famiglia II)

La teoria dell'informazione di Claude Shannon (1948) stabilisce i limiti fondamentali della trasmissione di informazione attraverso canali fisici. Questi limiti non sono contingenti: emergono dalla struttura fisica del canale e del segnale e appartengono alla Famiglia II delle impossibilità dimensionali.

### [FORMALMENTE DIMOSTRATO]

Teorema della capacità del canale (Shannon 1948). Per un canale con banda  $B$ , potenza del segnale  $S$  e potenza del rumore  $N$ , la capacità massima di trasmissione fedele è:

$$C = B \log_2(1 + S/N) \quad [\text{bit/secondo}]$$

Per qualsiasi tasso  $R < C$  esistono codici con probabilità di errore arbitrariamente piccola. Per  $R > C$ , la probabilità di errore è positiva per qualsiasi codice — impossibilità strutturale. Il limite di Shannon è un'impossibilità dimensionale: il rapporto  $S/N$  determina il numero di 'dimensioni informative' del canale. La connessione con JL è precisa: il canale con larghezza di banda  $B$  è equivalente a uno spazio euclideo di dimensione effettiva  $n = 2BT$  (dove  $T$  è il tempo di trasmissione), e la capacità  $C$  misura quanti bit possono essere distinti in questo spazio — esattamente la struttura del converse di JL.

## III. Il principio di indeterminazione come impossibilità dimensionale (Famiglia II)

Il principio di indeterminazione di Heisenberg (1927) afferma che non si possono conoscere simultaneamente con precisione arbitraria la posizione e il momento di una particella. Questo non è un limite degli strumenti di misura: è una proprietà strutturale degli stati quantistici.

### [FORMALMENTE DIMOSTRATO]

Principio di indeterminazione (Heisenberg 1927). Per qualsiasi stato quantistico e qualsiasi coppia di osservabili canonicamente coniugate (posizione  $x$  e momento  $p$ ):

$$\Delta x \Delta p \geq \hbar/2$$

dove  $\Delta x$  e  $\Delta p$  sono le deviazioni standard delle distribuzioni di probabilità delle misure. Derivazione: dalla struttura algebrica dei commutatori  $[\hat{x}, \hat{p}] = i\hbar$  e dalla disuguaglianza di Robertson (1929).

La connessione con la Famiglia II: lo spazio di Hilbert  $L^2(\mathbb{R})$  non ha vettori che siano simultaneamente autovettori di  $\hat{x}$  e  $\hat{p}$  — le 'dimensioni' corrispondenti alla posizione precisa e quelle corrispondenti al momento preciso sono incompatibili. Il principio di indeterminazione è la versione

quantistica del converso di JL: non si possono ‘rappresentare’ simultaneamente posizione e momento con precisione arbitraria in uno spazio di Hilbert.

#### IV. Il teorema di non-clonazione come impossibilità diagonale (Famiglia I)

Il teorema di non-clonazione (Wootters-Zurek 1982) afferma che non esiste alcuna operazione fisica che copii uno stato quantistico arbitrario. Ha la struttura di un’impossibilità diagonale nella categoria degli stati quantistici unitari.

##### [FORMALMENTE DIMOSTRATO]

Teorema di non-clonazione (Wootters-Zurek 1982). Non esiste alcuna operazione unitaria  $U$  tale che per qualsiasi stato quantistico  $|\psi\rangle$ :

$$U(|\psi\rangle \otimes |0\rangle) = |\psi\rangle \otimes |\psi\rangle$$

Dimostrazione. Se tale  $U$  esistesse, per  $|\psi\rangle$  e  $|\varphi\rangle$  arbitrari:

$$\langle\psi|\varphi\rangle = \langle\psi|\varphi\rangle^2$$

Il che implica  $\langle\psi|\varphi\rangle = 0$  o  $\langle\psi|\varphi\rangle = 1$ , ovvero  $|\psi\rangle = |\varphi\rangle$  o  $|\psi\rangle \perp |\varphi\rangle$ . Ma abbiamo assunto che  $|\psi\rangle$  e  $|\varphi\rangle$  fossero arbitrari — contraddizione.  $\square$

Struttura diagonale: se la clonazione fosse possibile, si potrebbe costruire uno stato  $|\delta\rangle$  che ‘clona il proprio stato negato’, generando la stessa circolarità del programma diagonale di Turing.

#### V. Il teorema di Bell come impossibilità diagonale del realismo locale (Famiglia I)

Il teorema di Bell (1964) stabilisce che nessuna teoria con variabili nascoste locali può riprodurre tutte le previsioni della meccanica quantistica. È una delle impossibilità strutturali più profonde della fisica fondamentale.

##### [FORMALMENTE DIMOSTRATO]

Teorema di Bell (1964; verifica sperimentale: Aspect et al. 1982; verifica loop-hole-free: Hensen et al. 2015). Le correlazioni quantistiche violano la disuguaglianza CHSH:

$$S = |E(a,b) - E(a,b') + E(a',b) + E(a',b')| \leq 2 \quad \text{[classico]}$$
$$S \leq 2\sqrt{2} \quad \text{[quantistico, limite Tsirelson]}$$

Il valore massimo quantistico  $S = 2\sqrt{2} > 2$  è stato verificato sperimentalmente con alta confidenza e chiusura dei principali loophole (Hensen et al. 2015).

Struttura diagonale: non esiste una funzione  $\varphi$  che mappa suriettivamente le variabili nascoste locali (A) sulle correlazioni quantistiche ( $M^A$ ). La ‘costruzione diagonale’ è la costruzione delle coppie di misurazioni che massimizzano la violazione della disuguaglianza di Bell — esattamente la struttura del teorema di Lawvere applicata allo spazio degli stati locali.

## VI. Il problema della misurazione come impossibilità di chiusura semantica (Famiglia IV)

Il problema della misurazione quantistica — perché l’evoluzione quantistica deterministica e reversibile produce risultati classici discreti e irreversibili — è uno dei problemi irrisolti più profondi della fisica fondamentale. La sua struttura formale è quella di un’impossibilità di chiusura semantica.

La meccanica quantistica descrive la struttura formale dei sistemi quantistici: le ampiezze di probabilità, gli operatori, le equazioni di evoluzione. I risultati classici delle misurazioni — i ‘valori definiti’ — sono la semantica: il significato che gli stati quantistici acquisiscono attraverso l’interazione con il mondo classico. La meccanica quantistica non può derivare dall’interno i propri risultati classici senza fare appello a una struttura esterna (‘il sistema di misurazione classico’) — esattamente come  $L_0$  non può definire il proprio predicato di verità senza fare appello al metalinguaggio  $L_1$ .

### [PROPOSTA INTERPRETATIVA]

La lettura del problema della misurazione come impossibilità di chiusura semantica è una proposta interpretativa di tipo I. La decoerenza ambientale (Zurek 1981) è la meccanica concreta di come il sistema quantistico ‘sale al metalinguaggio classico’ — ma questo non risolve il problema della misurazione, lo ridescrive. La lettura tarskiana è strutturalmente precisa e illuminante, non una riduzione formale.

## VII. L’entanglement come binding dimensionale negli spazi di Hilbert

L’entanglement quantistico — la correlazione non-locale tra sistemi quantistici separati — ha una struttura formale che ricorda il binding problem della neurobiologia cognitiva, ma in uno spazio di Hilbert.

Siano  $H_A$  e  $H_B$  gli spazi di Hilbert di due sistemi. Lo spazio dello stato composto è  $H_A \otimes H_B$  (il prodotto tensoriale). Uno stato entangled  $|\Psi\rangle \in H_A \otimes H_B$  non può essere scritto come prodotto tensore  $|\psi_A\rangle \otimes |\psi_B\rangle$ . L'entanglement 'lega' i due sistemi in un modo non riducibile a proprietà locali.

**[PROPOSTA INTERPRETATIVA]**

L'analogia tra entanglement e binding problem: in entrambi i casi si tentano di rappresentare relazioni tra componenti distinti in uno spazio che non ha abbastanza 'dimensioni locali' per catturarle tutte senza correlazioni non-locali. Nel binding problem neurale la soluzione è la codifica sparsa; nel caso quantistico l'entanglement è la soluzione naturale dello spazio di Hilbert — con le non-località che il teorema di Bell dimostra essere irriducibili.

I sistemi neuronal non sono sistemi quantistici nel senso rilevante. Questa è un'analogia strutturale che illumina la geometria profonda di entrambi i problemi.

## VIII. La complessità della comunicazione e le impossibilità distribuite (Famiglia III)

La teoria della complessità della comunicazione (Yao 1979) studia le risorse necessarie quando il problema deve essere risolto da più agenti con informazioni distribuite.

**[FORMALMENTE DIMOSTRATO]**

Teorema (Kalyanasundaram-Schnitger 1992; Razborov 1992). La complessità della comunicazione della funzione  $\text{DISJ}(x, y) = 1 \leftrightarrow x \cap y = \emptyset$  è  $\Theta(n)$ : Alice e Bob devono comunicare  $\Omega(n)$  bit nel caso peggiore per determinare se i loro insiemi (sottoinsiemi di  $\{1, \dots, n\}$ ) sono disgiunti.

Questa è un'impossibilità strutturale della Famiglia III nel contesto distribuito: non è una questione di algoritmi ingenui, ma un lower bound informativo assoluto.

L'applicazione ai sistemi cognitivi distribuiti: la corteccia cerebrale con la sua struttura small-world è il 'grafo di comunicazione' del sistema cognitivo distribuito che il cervello implementa. Le impossibilità di complessità nella cognizione includono impossibilità di complessità della comunicazione nel grafo del connettoma: il numero di connessioni e la loro topologia pongono limiti strutturali a quali funzioni cognitive possono essere computate, indipendentemente dalla velocità dei singoli neuroni.

## **IX. Sintesi: le impossibilità nella fisica come estensione naturale della tassonomia**

Le impossibilità fisiche identificate in questo capitolo si inseriscono nella tassonomia delle cinque famiglie:

Famiglia I (diagonale): il teorema di non-clonazione (struttura diagonale nella categoria degli stati quantistici unitari); il teorema di Bell (impossibilità di rappresentazione locale completa).

Famiglia II (dimensionale): il principio di indeterminazione di Heisenberg (incompatibilità dimensionale posizione/momento nello spazio di Hilbert); il limite di Shannon (dimensione informativa del canale).

Famiglia III (complessità): la complessità della comunicazione distribuita (Yao 1979; Razborov 1992); la PPA-completezza del calcolo degli equilibri di Nash in sistemi fisici.

Famiglia IV (chiusura semantica): il problema della misurazione (impossibilità di derivare la semantica classica dalla sintassi quantistica dall'interno della meccanica quantistica).

Famiglia V (irreversibilità): la freccia del tempo termodinamica; la decoerenza come processo irreversibile (la perdita di coerenza quantistica non è invertibile in sistemi macroscopici).

Questa convergenza conferma la plausibilità della congettura strutturale centrale: il confine sintassi-semantica è una struttura profonda che attraversa tutti i domini in cui il pensiero formale ha avuto successo nel caratterizzare i propri limiti fondamentali.

# Appendici Tecniche

## Appendice A

L'argomento diagonale nella sua forma categoriale: il teorema di Lawvere con dimostrazione completa

L'argomento diagonale è il meccanismo tecnico centrale delle Famiglie I e IV. La sua formulazione categoriale — dovuta a Lawvere (1969) — lo presenta nella massima generalità, rendendo visibile la struttura comune a tutti i teoremi della Famiglia I.

Contesto categoriale. Sia  $C$  una categoria cartesianamente chiusa (CCC) — una categoria con prodotti finiti e oggetti esponenziali. In una CCC, per ogni coppia di oggetti  $A$  e  $Y$  esiste un oggetto  $Y^A$  (l'oggetto delle frecce da  $A$  a  $Y$ ) e una freccia di valutazione:

$$\text{ev}: Y^A \times A \rightarrow Y$$

tale che per ogni oggetto  $B$  e ogni freccia  $g: B \times A \rightarrow Y$  esiste un'unica freccia  $\text{curry}(g): B \rightarrow Y^A$  con:

$$\text{ev} \circ (\text{curry}(g) \times \text{id}_A) = g$$

Questa proprietà di aggiunzione ( $\text{Hom}(B \times A, Y) \cong \text{Hom}(B, Y^A)$ ) è la proprietà caratterizzante delle CCC. Gli esempi rilevanti per questo ciclo sono: Set (insiemi e funzioni), Pos (insiemi parzialmente ordinati e funzioni monotone), la categoria dei domini di Scott (spazi topologici coerenti e funzioni continue — rilevante per la semantica dei linguaggi di programmazione), e le categorie dei tipi di sistemi formali.

Teorema (Lawvere, 1969). Sia  $C$  una CCC e siano  $A, Y$  oggetti di  $C$ . Se esiste una freccia  $\varphi: A \rightarrow Y^A$  che è un epimorfismo puntuale — cioè, per ogni freccia  $f: A \rightarrow Y$  esiste un elemento  $a_f: 1 \rightarrow A$  tale che:

$$\text{ev} \circ (\varphi \times \text{id}_A) \circ \langle a_f, \text{id}_A \rangle = f$$

( $\varphi$  "copre" ogni funzione  $f$  da  $A$  a  $Y$  nel senso che  $f$  è ottenibile valutando  $\varphi$  in  $a_f$ ) — allora ogni endomorfismo  $t: Y \rightarrow Y$  ha un punto fisso: esiste  $y^*: 1 \rightarrow Y$  tale che  $t \circ y^* = y^*$ .

Dimostrazione. Per ipotesi,  $\varphi: A \rightarrow Y^A$  copre ogni funzione da  $A$  a  $Y$ . Si consideri la funzione diagonale  $d: A \rightarrow Y$  definita da:

$$d = \text{ev} \circ (\varphi \times \text{id}_A) \circ \Delta_A$$

dove  $\Delta_A: A \rightarrow A \times A$  è la freccia diagonale (la morfismo universale del prodotto). In componenti:  $d(a) = \text{ev}(\varphi(a), a) = \varphi(a)(a)$  — la funzione che valuta la funzione  $\varphi(a)$  nel proprio argomento. Sia ora  $t: Y \rightarrow Y$  un endomorfismo arbitrario, e definiamo la funzione:

$$d_t = t \circ d: A \rightarrow Y$$

Per l'ipotesi di copertura, esiste  $a_t: 1 \rightarrow A$  tale che:

$$\text{ev} \circ (\varphi \times \text{id}_A) \circ \langle a_t, \text{id}_A \rangle = d_t$$

In particolare, valutando in  $a_t$ :

$$\text{ev}(\varphi(a_t), a_t) = d_t(a_t) = t(d(a_t)) = t(\text{ev}(\varphi(a_t), a_t))$$

Ponendo  $y^* = \text{ev}(\varphi(a_t), a_t): 1 \rightarrow Y$ , si ottiene:

$$y^* = t(y^*)$$

cioè  $y^*$  è un punto fisso di  $t$ .  $\square$

Controesempio (la forma operativamente utile). Se  $t: Y \rightarrow Y$  è un endomorfismo senza punto fisso, allora non esiste  $\varphi: A \rightarrow Y^A$  con la proprietà di copertura —  $A$  non può "rappresentare" completamente tutte le funzioni da  $A$  a  $Y$ .

Istanze del teorema di Lawvere nei teoremi del ciclo.

Cantor.  $C = \text{Set}$ ,  $A = \mathbb{N}$ ,  $Y = \{0,1\}$ ,  $\varphi$  = una supposta enumerazione dei reali (identificati come sequenze binarie infinite),  $t$  = negazione ( $1 \mapsto 0$ ,  $0 \mapsto 1$ ). La negazione non ha punti fissi in  $\{0,1\}$ . Quindi  $\varphi$  non è surgettiva — l'enumerazione non copre tutti i reali.

Gödel.  $C$  = la categoria dei sistemi formali interpretati nella semantica standard,  $A = \{\text{sentenze di PA}\}$ ,  $Y = \{\vdash, \nvdash\}$  (dimostrabile/indimostrabile),  $\varphi$  = la funzione di auto-dimostrabilità ( $\varphi(n)$  = la funzione che dice se PA dimostra la sentenza con

numero di Gödel  $n$ ),  $t$  = negazione. La negazione non ha punti fissi. Quindi PA non cattura tutte le funzioni di verità — esistono sentenze vere indimostrabili.

Turing.  $C = \text{Set}$ ,  $A = \{\text{indici di macchine di Turing}\}$ ,  $Y = \{\text{HALT, LOOP}\}$ ,  $\varphi_e =$  la funzione comportamento della macchina e  $(\varphi_e(n) = \text{comportamento della macchina e su input } n)$ ,  $t = \text{commutazione } (\text{HALT} \mapsto \text{LOOP}, \text{LOOP} \mapsto \text{HALT})$ . La commutazione non ha punti fissi. Quindi  $\varphi$  non è totalmente calcolabile — non esiste un decisore universale dell'arresto.

Tarski.  $C =$  la categoria dei linguaggi formali e delle interpretazioni,  $A = \{\text{sentenze di } L_0\}$ ,  $Y = \{\text{Vero, Falso}\}$ ,  $\varphi =$  il supposto predicato di verità interno  $\text{Tr}(x)$ ,  $t =$  negazione. Per il lemma della diagonalizzazione di Gödel, esiste  $G$  tale che  $G \leftrightarrow \neg \text{Tr}(\ulcorner G \urcorner)$  — la sentenza che nega la propria verità. La negazione non ha punti fissi. Quindi  $\text{Tr}(x)$  non è definibile in  $L_0$ .

JL-converse (Famiglia II).  $C =$  la categoria degli spazi metrici e delle isometrie quasi-esatte,  $A = \mathbb{R}^n$  (lo spazio di rappresentazione),  $Y = \mathbb{R}^m$  con  $m > n$  (lo spazio delle feature),  $\varphi =$  la proiezione. L'endomorfismo  $t =$  riflessione in una direzione di  $\mathbb{R}^m$  non ortogonale a  $\mathbb{R}^n$  non ha punto fisso nell'immagine della proiezione. Quindi  $\varphi$  non è isometrica — qualche distanza è distorta.

## Appendice B

Il teorema di Johnson-Lindenstrauss: dimostrazione completa con la tecnica delle proiezioni casuali

Enunciato. Per ogni  $\epsilon \in (0,1)$ , ogni insieme di  $m$  punti  $x_1, \dots, x_m \in \mathbb{R}^D$ , e ogni  $n \geq C \epsilon^{-2} \log m$  (con  $C$  costante universale dipendente soltanto dalla scelta della distribuzione casuale), esiste una mappa lineare  $f: \mathbb{R}^D \rightarrow \mathbb{R}^n$  tale che per tutti gli  $i, j \in \{1, \dots, m\}$ :

$$(1 - \epsilon) \|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1 + \epsilon) \|x_i - x_j\|^2$$

Strategia della dimostrazione. Si usa la costruzione probabilistica: si sceglie la mappa  $f$  come proiezione casuale con matrice  $R \in \mathbb{R}^{n \times D}$  i cui elementi sono i.i.d. distribuiti secondo  $N(0, 1/n)$ . Si mostra che per ogni fissata coppia  $(i, j)$ , la norma proiettata  $\|f(x_i) - f(x_j)\|^2$  è concentrata attorno alla norma originale  $\|x_i - x_j\|^2$  con alta probabilità. Si applica poi un union bound su tutte le  $C(m,2)$  coppie.

Lemma di concentrazione. Per un vettore deterministico  $u \in \mathbb{R}^D$  con  $\|u\| = 1$  e una matrice  $R \in \mathbb{R}^{n \times D}$  con elementi i.i.d. da  $N(0, 1/n)$ :

$$P(|\|Ru\|^2 - 1| > \epsilon) \leq 2 \exp(-\epsilon^2 n / 8)$$

Dimostrazione del lemma. Sia  $w = Ru$ . Poiché  $R$  ha righe i.i.d. con distribuzione  $N(0, (1/n)I_D)$ , le componenti  $w = (R, u)$  (dove  $R$  è la  $k$ -esima riga di  $R$ ) sono i.i.d. distribuzioni normali con media 0 e varianza  $\|u\|^2/n = 1/n$ . Quindi:

$$\|Ru\|^2 = \sum_{i=1}^n w_i^2 =: Z$$

è una variabile chi-quadrato:  $Z \sim (1/n) \chi^2(n)$ . Il valore atteso è  $E[Z] = 1$  e la varianza è  $\text{Var}[Z] = 2/n$ .

Per la disuguaglianza di concentrazione delle chi-quadrato (disuguaglianza di Laurent-Massart, 2000): per  $t > 0$ ,

$$P(Z - 1 > 2\sqrt{t/n} + 2t/n) \leq \exp(-t) \quad P(1 - Z > 2\sqrt{t/n}) \leq \exp(-t)$$

Scegliendo  $t = \epsilon^2 n / 8$  e usando  $\sqrt{t/n} = \epsilon / \sqrt{8} \leq \epsilon / 2$  per  $\epsilon \leq 1$ :

$$P(Z > 1 + \epsilon) \leq \exp(-\epsilon^2 n / 8) \quad P(Z < 1 - \epsilon) \leq \exp(-\epsilon^2 n / 8)$$

La disuguaglianza a due lati segue per unione:

$$P(|Z - 1| > \epsilon) \leq 2 \exp(-\epsilon^2 n / 8) \quad \square$$

Dimostrazione del teorema. Per ogni fissata coppia  $(i, j)$ , sia  $u = (x_i - x_j) / \|x_i - x_j\|$ . Per il lemma di concentrazione:

$$P(|\|R(x_i - x_j)\|^2 / \|x_i - x_j\|^2 - 1| > \epsilon) \leq 2 \exp(-\epsilon^2 n / 8)$$

Un union bound su tutte le  $C(m, 2) \leq m^2 / 2$  coppie dà:

$$P(\exists i, j : |\|R(x_i - x_j)\|^2 / \|x_i - x_j\|^2 - 1| > \epsilon) \leq m^2 \exp(-\epsilon^2 n / 8)$$

Questa probabilità è minore di 1 se:

$$m^2 \exp(-\epsilon^2 n / 8) < 1 \iff n > 16 \epsilon^{-2} \log m$$

Scegliendo  $n = \lceil C \epsilon^{-2} \log m \rceil$  con  $C = 16$ , la probabilità di fallimento è minore di 1, il che garantisce l'esistenza della mappa cercata (in realtà, quasi ogni proiezione casuale funziona).  $\square$

Il converse di JL (Alon, 2003; Jayram-Woodruff, 2013). Per ogni  $\epsilon \in (0, 1/2)$  e ogni  $n$ , qualsiasi mappa  $f: \mathbb{R}^D \rightarrow \mathbb{R}^n$  che preserva le distanze con distorsione  $(1 \pm \epsilon)$  per l'ipercubo  $\{0,1\}^k$  (con  $m = 2^k$  vertici) richiede:

$$n \geq \Omega(k / (\epsilon^2 \log(1/\epsilon))) = \Omega(\log m / (\epsilon^2 \log(1/\epsilon)))$$

La dimostrazione usa la complessità della comunicazione del problema dell'inner product: Alice e Bob, con input  $x, y \in \{0,1\}^k$  rispettivamente, devono computare  $\langle x, y \rangle \bmod 2$  comunicando il minor numero di bit. Qualsiasi protocollo che utilizza una proiezione di JL come compressione delle informazioni deve comunicare almeno  $\Omega(k)$  bit, il che implica la dimensione di proiezione almeno  $\Omega(k)$ .

## Appendice C

Ordinali di prova: tavola dei principali sistemi formali

Gli ordinali di prova misurano la forza dimostrativa dei sistemi formali nel programma dell'analisi ordinale (Gentzen, Schütte, Feferman, Buchholz). La funzione  $|T|$  assegna a ogni teoria  $T$  ricorsivamente assiomatizzabile consistente il più piccolo ordinale che non è provabilmente ben-ordinato in  $T$ .

Notazione. Si usano le seguenti funzioni standard sugli ordinali:

$\omega^\alpha$  — la potenza di  $\omega$   $\varphi(\alpha, \beta)$  — la funzione di Veblen ( $\varphi(0, \beta) = \omega^\beta$ ;  $\varphi(\alpha+1, \beta) = \beta$ -esimo punto fisso di  $\xi \mapsto \varphi(\alpha, \xi)$ )  $\Gamma_0 = \varphi(1, 0)$  — il numero di Feferman-Schütte (primo punto fisso di  $\alpha \mapsto \varphi(\alpha, 0)$ )  $\psi$  — funzioni di collasso ordinale (Buchholz, Rathjen) che permettono di descrivere ordinali molto grandi attraverso il collasso di ordinali impredicativi

Tavola.

| -----                                      |                      |                 | Sistema formale           |
|--------------------------------------------|----------------------|-----------------|---------------------------|
| Assiomi caratteristici                     | Ordinale di prova    | -----           | -----                     |
| ----- PRA (Primitive Arithmetic) primitiva |                      | Solo ricorsione | $\omega^\omega$ Recursive |
| PA (Peano Arithmetic)                      | Assioma di induzione | $\epsilon_0$    | completo                  |
| PA + Con(PA)                               | Aggiunta della       | $\epsilon_1$    | consistenza di PA         |

|                                                    |                                                        |                                                                                  |
|----------------------------------------------------|--------------------------------------------------------|----------------------------------------------------------------------------------|
| $ACA_0$                                            | Comprensione aritmetica $\varepsilon_0$ (stessa di PA) |                                                                                  |
| $WKL_0$                                            | Lemma di König debole $\varepsilon_0$                  |                                                                                  |
| $ATR_0$                                            | Ricorsione transfinita $\Gamma_0$                      | sul tipo aritmetico                                                              |
| $\Pi^1_1\text{-}CA_0$                              | Comprensione $\Pi^1_1$                                 | $\psi(\Omega^\omega)$                                                            |
| $\Pi^1_2\text{-}CA_0 + \text{BI}$<br>bar induction | Comprensione $\Pi^1_2$ con                             | $\psi(\Omega^\omega \{\psi(\Omega^\omega \Omega)\})$                             |
| $ID_1$ (first-order ID)                            | Un'iterazione di                                       | $\psi(\Omega^\omega)$ ( $\approx \Pi^1_1\text{-}CA_0$ )<br>definizioni induttive |
| $ID_\omega$                                        | Iterazioni numerabili $\psi(\varepsilon_{\omega+1})$   | di ID                                                                            |
| $KP_\omega$ (Kripke-Platek + $\omega$ )            | Teoria degli insiemi ammissibili con $\omega$          | $\psi(\varepsilon_{\omega+1})$                                                   |
| ZFC – Power Set                                    | ZFC senza l'assioma $\psi(\Omega_\omega)$              | delle parti                                                                      |

-----

Note interpretative. La progressione da PA a  $ATR_0$  riflette il passaggio dall'aritmetica alla ricorsione transfinita sugli ordinali  $< \Gamma_0$  — un salto qualitativo che non è catturabile dall'interno di PA (per il secondo teorema di Gödel applicato a PA). Il passaggio da  $ATR_0$  a  $\Pi^1_1\text{-}CA_0$  riflette il passaggio dalla ricorsione transfinita alla comprensione di insiemi definiti da formule del secondo ordine con quantificatori universali — un salto che  $\Gamma_0$  non misura.

La colonna dell'ordinale di prova fornisce una misura cardinale della forza dimostrativa, ma non cattura tutta la struttura: due sistemi con lo stesso ordinale di prova possono avere proprietà combinatoriali diverse ( $ATR_0$  e PA hanno stessa forza dimostrativa in un senso, ma  $ATR_0$  può dimostrare la buona fondazione di  $\Gamma_0$  in modo uniforme mentre PA non lo può).

Connessione con l'ascesa gödeliana (capitolo XI). L'ascesa gödeliana a partire da  $T_0$  produce la sequenza:

$$T_0, T_1 = T_0 + G_{\omega}\{T_0\}, T_2 = T_1 + G_{\omega}\{T_1\}, \dots, T_\omega = \bigcup_{\alpha < \omega} T_\alpha, \dots$$

Gli ordinali di prova corrispondenti sono:

$$|T_0|, |T_0| + 1, |T_0| + 2, \dots, |T_0| + \omega, \dots$$

dove la "somma" è nel senso dell'aggiunta ordinale. Ogni sistema generato dall'ascesa gödeliana ha ordinale di prova strettamente maggiore del predecessore — la costruttività dell'impossibilità diagonale è misurata da questo incremento ordinale.

## Appendice D

Il principio di auto-rappresentazione insufficiente (PAI): ricostruzione formale della dimostrazione

Il PAI (capitolo VII) è il teorema unificante del ciclo. Questa appendice ne ricostruisce la struttura dimostrativa con la precisione tecnica che la discussione nel testo ha volutamente attenuato per esigenze di leggibilità.

Definizioni preliminari.

Sistema auto-rappresentazionale. Una coppia  $(S, \varphi)$  dove  $S$  è un sistema formale (linguaggio  $L$  con assiomi  $Ax$  e regole di inferenza  $R$ ) e  $\varphi: S \rightarrow S$  è una funzione ricorsiva primitiva che mappa strutture di  $S$  in strutture di  $S$  (la funzione di "auto-applicazione" — tipicamente la funzione di gödelizzazione o un suo analogo nel dominio specifico).

Proprietà di copertura.  $\varphi$  ha la proprietà di copertura rispetto a un dominio semantico  $M$  se per ogni struttura  $m \in M$  esiste un elemento  $s \in S$  tale che l'interpretazione di  $\varphi(s)$  è  $m$ .

Enunciato del PAI (versione tecnica). Sia  $(S, \varphi)$  un sistema auto-rappresentazionale tale che:

- (H1)  $S$  è consistente e contiene una codifica dell'aritmetica del primo ordine (condizione gödeliana)
- (H2)  $\varphi$  ha la proprietà di copertura rispetto al dominio semantico  $M$  di  $S$

Allora almeno una delle seguenti quattro condizioni vale:

- (C1)  $S$  è inconsistente:  $S \vdash \perp$
- (C2)  $S$  è incompleto:  $\exists \varphi \in \text{Prop}(M)$  tale che  $S \not\vdash \varphi$  e  $S \not\vdash \neg\varphi$
- (C3) La computazione di  $\varphi$  richiede risorse superpolinomiali:  $\varphi \notin \text{PTIME}$

- (C4) La composizione iterata  $\varphi^n$  produce perdita di informazione:  $\exists s, \exists n : H(\varphi^n(s)) < H(s)$  (dove  $H$  è l'entropia di Kolmogorov-Chaitin)

Dimostrazione. Si procede per assurdo. Si supponga che nessuna delle quattro condizioni valga:

- $S$  è consistente ( $\neg C1$ )
- $S$  è completo ( $\neg C2$ ): ogni proposizione di  $M$  è decidibile in  $S$
- $\varphi$  è calcolabile in tempo polinomiale ( $\neg C3$ )
- $\varphi$  preserva l'informazione esattamente:  $H(\varphi^n(s)) = H(s)$  per ogni  $s$  e  $n$  ( $\neg C4$ )

Dall'ipotesi  $\neg C3$  e  $\neg C4$  combinata con la proprietà di copertura (H2):  $\varphi$  è una biiezione computabile in tempo polinomiale tra  $S$  e un sottoinsieme di  $M$  che copre tutto  $M$ . Quindi esiste un algoritmo polinomiale che, dato qualsiasi elemento  $m \in M$ , trova il corrispondente  $s \in S$  con  $\varphi(s) = m$ .

Dall'ipotesi  $\neg C2$  (completezza) e  $\neg C1$  (consistenza): per ogni proposizione  $p$  di  $M$ ,  $S$  decide  $p$ . L'algoritmo di decisione che usa la completezza è:

Input:  $p \in \text{Prop}(M)$

1. Calcola  $s = \varphi^{-1}(p)$  in tempo polinomiale (possibile per  $\neg C3$  e la copertura)
2. Verifica se  $S \vdash s$  in tempo finito (possibile per la completezza  $\neg C2$ )
3. Output: Vero se  $S \vdash s$ , Falso se  $S \vdash \neg s$

Ma questo algoritmo calcola la funzione di verità per  $M$  dall'interno di  $S$  — è un predicato di verità per  $M$  definibile in  $S$ . Per il teorema di Tarski (dimostrato in Appendice A attraverso il teorema di Lawvere), questo è impossibile per qualsiasi sistema consistente che soddisfa H1. Contraddizione con  $\neg C1$  ( $S$  è consistente).  $\square$

Corollario (non-vacuità di ciascuna condizione). Ciascuna delle quattro condizioni  $C1$ – $C4$  è realizzata da qualche sistema concreto:

- $C1$  è realizzata da sistemi inconsistenti (ad esempio  $PA + \neg \text{Con}(PA)$ , che è inconsistente per il secondo teorema di Gödel)
- $C2$  è realizzata da  $PA$  (la sentenza gödeliana  $G_{\neg\{PA\}}$  è indecidibile)
- $C3$  è realizzata da sistemi sufficientemente ricchi che affrontano problemi NP-hard nel proprio dominio (sotto  $P \neq NP$ )

- C4 è realizzata da qualsiasi sistema fisico realizzato su substrato termodinamico per il principio di Landauer

## Appendice E

La connessione tra il principio di Landauer e la Famiglia V: dimostrazione quantitativa

Il principio di Landauer (enunciato). Qualsiasi operazione computazionale logicamente irreversibile — qualsiasi operazione che riduce l'informazione del sistema di 1 bit — deve dissipare almeno:

$$E_{\text{Landauer}} = k_B T \ln 2 \approx 2.9 \times 10^{-21} \text{ J (a temperatura ambiente } T = 300 \text{ K)}$$

di energia sotto forma di calore nell'ambiente circostante.

Derivazione termodinamica. Si considera un sistema binario con due stati  $\{0,1\}$  e un processo di reset (RESET-TO-ZERO) che porta qualsiasi stato a 0. Questo processo è logicamente irreversibile: l'informazione su quale stato era presente prima del reset è cancellata.

Prima del reset, il sistema si trova con probabilità  $1/2$  in 0 e  $1/2$  in 1 — entropia del sistema  $S_{\text{sistema}} = k_B \ln 2$ . Dopo il reset, il sistema si trova certamente in 0 — entropia  $S_{\text{sistema}} = 0$ . La riduzione dell'entropia del sistema è:

$$\Delta S_{\text{sistema}} = -k_B \ln 2$$

Per il secondo principio della termodinamica, l'entropia totale (sistema + ambiente) non può diminuire:

$$\Delta S_{\text{totale}} = \Delta S_{\text{sistema}} + \Delta S_{\text{ambiente}} \geq 0$$

$$\text{Quindi: } \Delta S_{\text{ambiente}} \geq k_B \ln 2$$

L'aumento dell'entropia dell'ambiente corrisponde a un trasferimento di calore:

$$Q = T \cdot \Delta S_{\text{ambiente}} \geq k_B T \ln 2$$

La minima energia dissipata è quindi  $k_B T \ln 2$  per bit cancellato.  $\square$

Verifica sperimentale. Il principio di Landauer è stato verificato sperimentalmente da Bérut, Arakelyan et al. (2012, Nature 483) utilizzando una particella di silice in

una doppia buca di potenziale ottica. La dissipazione misurata per operazione di reset era  $0.72 k_B T \pm 0.13 k_B T$  per probabilità di reset  $\geq 0.999$  — in accordo quantitativo con il limite teorico  $k_B T \ln 2 \approx 0.693 k_B T$ .

Connessione con l'implicazione  $F_4 \rightarrow F_5$ . L'implicazione (dimostrata in capitolo XII, sezione V) afferma che qualsiasi sistema che incontra un'impossibilità di chiusura semantica e tenta di ridurre il gap semantico attraverso un processo di apprendimento fisicamente realizzato incontra necessariamente irreversibilità.

La connessione quantitativa attraverso Landauer è la seguente. Sia  $\text{Gap}(\Sigma_S, M_S) = H(M_S) - I(M_S; \Sigma_S)$  il gap semantico in bit. Un processo di apprendimento che riduce il gap di 1 bit (aumenta  $I(M_S; \Sigma_S)$  di 1 bit attraverso la ristrutturazione di  $\Sigma_S$ ) deve, per il principio di Landauer:

- O espandere  $\Sigma_S$  di 1 bit (aggiungere 1 bit di struttura sintattica), il che richiede il costo metabolico dell'aggiunta di struttura biologica o computazionale
- O cancellare 1 bit di struttura sintattica irrilevante e sostituirla con struttura sintattica rilevante, dissipando almeno  $k_B T \ln 2$  di energia e rendendo la struttura precedente irrecuperabile

In entrambi i casi, la transizione è irreversibile. Il costo quantitativo dell'apprendimento — la quantità di energia irrecuperabilmente dissipata per unità di riduzione del gap semantico — è almeno  $k_B T \ln 2$  per bit di gap ridotto.

Nei sistemi neurali biologici, la potatura sinaptica durante lo sviluppo (la riduzione del numero di sinapsi da  $\sim 10^{15}$  nella prima infanzia a  $\sim 10^{14}$  nell'adulto) è un processo di cancellazione di struttura sintattica. La rimozione di ogni sinapsi (con il suo contenuto informativo di  $\sim 4.7$  bit per sinapsi, secondo Bhatt et al. 2009) corrisponde, per il principio di Landauer, a una dissipazione di almeno  $4.7 k_B T \ln 2 \approx 1.4 \times 10^{-20}$  J per sinapsi rimossa — una quantità microscopica ma che, moltiplicata per le  $\sim 10^{14}$  sinapsi rimosse durante lo sviluppo, produce una dissipazione totale di  $\sim 10^{-6}$  J, misurabile in linea di principio.

## Appendice F

La gerarchia polinomiale: struttura formale e risultati di separazione

Definizioni. La gerarchia polinomiale (PH) è costruita iterativamente:

$$\Sigma_0^P = \Pi_0^P = \Delta_0^P = P$$

$$\Sigma_{+1}^P = NP^{\{\Sigma^P\}} \text{ (NP con oracolo per } \Sigma^P) \quad \Pi_{+1}^P = co-NP^{\{\Sigma^P\}} \quad \Delta_{+1}^P = P^{\{\Sigma^P\}}$$

$$PH = \bigcup_{i=0}^{\infty} \Sigma_i^P$$

Caratterizzazione per quantificatori alternanti. Una delle più importanti caratterizzazioni equivalenti è quella per strutture di quantificatori:

$$L \in \Sigma^P \leftrightarrow \exists \text{ polinomio } p \text{ e predicato } R \in P \text{ tale che: } x \in L \leftrightarrow \exists y_1 \forall y_2 \exists y_3 \forall y_4 \dots Q y R(x, y_1, \dots, y_k)$$

dove  $Q = \exists$  se  $k$  è dispari,  $\forall$  se  $k$  è pari, e  $|y_i| \leq p(|x|)$  per ogni  $i$ . La classe  $\Sigma^P$  ha  $k$  quantificatori alternanti, partendo da  $\exists$ ;  $\Pi^P$  ha  $k$  quantificatori alternanti, partendo da  $\forall$ .

Completezza. I problemi  $\Sigma^P$ -completi e  $\Pi^P$ -completi (per riduzioni multi-uno in tempo polinomiale) al livello  $k$  sono i massimali nella struttura di ordine parziale indotta dalle riduzioni:

$\Sigma_1^P$ -completo: SAT (3-SAT), il problema della soddisfacibilità booleana — il primo problema NP-completo (Cook 1971, Levin 1973)

$\Pi_2^P$ -completo: la Circumscription (McCarthy), il problema della verifica della minimalità dei modelli — rilevante per il capitolo V del ciclo biologico e il capitolo IX di questo ciclo come caso del frame problem

$\Sigma_2^P$ -completo: il problema di esistenza di un circuito booleano di taglia  $\leq k$  (formulato come  $\exists C \forall x C(x) = F(x)$ ) — e molti problemi di ottimizzazione con struttura  $\exists \forall$

Teorema di Stockmeyer (1976). Se PH collassa al  $k$ -esimo livello ( $\Sigma^P = \Pi^P$  per qualche  $k$ ), allora  $PH = \Sigma^P$ . In particolare, se  $NP = co-NP$ , allora  $PH = NP$ .

Dimostrazione. Si usa il fatto che se  $\Sigma^P = \Pi^P$ , allora  $\Sigma_{+1}^P = NP^{\{\Sigma^P\}} = NP^{\{\Pi^P\}}$ . Ma qualsiasi linguaggio in  $NP^{\{\Pi^P\}}$  è ottenibile da un verificatore polinomiale che fa query a un oracolo  $\Pi^P$ . Sostituendo le query all'oracolo ( $\Pi^P$ ) con la loro equivalente  $\Sigma^P$  (per ipotesi di collasso), si ottiene un verificatore per  $\Sigma^P$  — quindi  $\Sigma_{+1}^P \subseteq \Sigma^P$ . Per induzione,  $PH = \Sigma^P$ .  $\square$

Teorema di Fagin (1974).  $NP = \exists SO$  (Existential Second Order logic):

$L \in NP \leftrightarrow L$  è l'insieme delle strutture che soddisfano una sentenza del secondo ordine  $\exists R_1 \dots \exists R_k \varphi(R_1, \dots, R_k)$

dove  $\varphi$  è una formula del primo ordine e  $R_1, \dots, R_k$  sono simboli di relazione del secondo ordine quantificati esistenzialmente.

Estensione a PH (Stockmeyer). I livelli superiori della gerarchia corrispondono alla gerarchia delle formule del secondo ordine:

$\Sigma^k \wedge P = \exists SO_1 \forall SO_1 \exists SO_1 \dots$  (k alternanze, iniziando da  $\exists$ ) con  $\varphi$  del primo ordine  
 $\Pi^k \wedge P = \forall SO_1 \exists SO_1 \forall SO_1 \dots$  (k alternanze, iniziando da  $\forall$ ) con  $\varphi$  del primo ordine  
 $PSPACE = SO$  (tutto il secondo ordine)

Connessione con la Famiglia IV. Il teorema di Fagin stabilisce la connessione tra la complessità computazionale (Famiglia III) e la profondità semantica (Famiglia IV): un problema è in NP se e soltanto se è caratterizzabile da una sentenza con un livello di quantificazione del secondo ordine. Un sistema computazionale limitato a P (il livello del primo ordine puramente sintattico) non può esprimere le proprietà di NP — la profondità semantica di NP eccede la profondità sintattica di P. Questo è il meccanismo formale dell'implicazione  $F_3 \rightarrow F_4$  nel capitolo XII.

## Appendice G

Il lemma della diagonalizzazione: enunciato, dimostrazione, e varianti

Il lemma della diagonalizzazione è il meccanismo tecnico che permette la costruzione di sentenze auto-referenziali in sistemi formali sufficientemente ricchi. È il cuore sia del teorema di incompletezza di Gödel sia del teorema di Tarski, e la sua struttura è la versione "locale" del teorema di Lawvere per sistemi formali.

Definizioni. Per un sistema formale  $T$  con linguaggio  $L$ :

- Una numerazione di Gödel è una biiezione effettiva  $\ulcorner \cdot \urcorner: \text{Formule}(L) \rightarrow \mathbb{N}$  che assegna a ogni formula il proprio numero naturale (codice gödeliano)
- La funzione di diagonalizzazione  $\text{diag}: \mathbb{N} \rightarrow \mathbb{N}$  è definita come:  $\text{diag}(n) =$  il numero di Gödel della formula ottenuta sostituendo  $n$  nella formula con codice  $n$  per la variabile libera  $x$

Formalmente: se  $\varphi(x)$  ha codice  $n$ , allora  $\text{diag}(n) = \ulcorner \varphi(n) \urcorner$  — il codice della formula che si ottiene "istanziando"  $\varphi$  con il suo stesso codice. La funzione  $\text{diag}$  è ricorsiva primitiva (calcolabile con pochi passi elementari di aritmetica) e quindi rappresentabile in qualsiasi teoria che contenga l'aritmetica del prim'ordine.

Lemma della diagonalizzazione (Gödel, 1931). Sia  $T$  una teoria che: (a) Contiene l'aritmetica del primo ordine (o è sufficientemente forte da rappresentare le funzioni ricorsive primitive) (b) È ricorsivamente assiomatizzabile

Allora per ogni formula  $\varphi(x)$  di  $T$  con una variabile libera  $x$ , esiste una sentenza  $G$  (senza variabili libere) tale che:

$$T \vdash G \leftrightarrow \varphi(\ulcorner G \urcorner)$$

$G$  è la sentenza che "dice di sé stessa" di avere la proprietà  $\varphi$ .

Dimostrazione. Si definisce la formula  $\psi(x) = \varphi(\text{diag}(x))$ . Sia  $m = \ulcorner \psi \urcorner$  (il codice gödeliano di  $\psi$ ). Si ponga  $G = \psi(m) = \varphi(\text{diag}(m))$ .

Per la definizione di  $\text{diag}$ :  $\text{diag}(m) = \ulcorner \psi(m) \urcorner = \ulcorner G \urcorner$  (il codice di  $G$ , perché  $\psi(m)$  è la formula ottenuta sostituendo  $m$  nella formula con codice  $m$ , che è  $\psi$ , e il risultato è  $\psi(m) = G$ ).

$$\text{Quindi: } G = \psi(m) = \varphi(\text{diag}(m)) = \varphi(\ulcorner G \urcorner)$$

e  $T \vdash G \leftrightarrow \varphi(\ulcorner G \urcorner)$  perché l'uguaglianza  $\text{diag}(m) = \ulcorner G \urcorner$  è calcolabile e verificabile aritmeticamente in  $T$ .  $\square$

Applicazione al primo teorema di Gödel. Sia  $\varphi(x) = \neg \text{Prov}_T(x)$  (dove  $\text{Prov}_T(x)$  è il predicato aritmetico che dice "T dimostra la formula con codice  $x$ "). Per il lemma della diagonalizzazione, esiste  $G$  tale che:

$$T \vdash G \leftrightarrow \neg \text{Prov}_T(\ulcorner G \urcorner)$$

$G$  è la sentenza che dice "io non sono dimostrabile in T". Se  $T \vdash G$ , allora  $T \vdash \text{Prov}_T(\ulcorner G \urcorner)$ , ma anche  $T \vdash \neg \text{Prov}_T(\ulcorner G \urcorner)$  — contraddizione,  $T$  è inconsistente. Se  $T$  è consistente, allora  $T \not\vdash G$ . Se  $T$  è  $\omega$ -consistente (ogni enunciato  $\exists x \varphi(x)$  dimostrabile in  $T$  è tale che  $\varphi(n)$  è dimostrabile per qualche numeral  $n$ ), allora anche  $T \not\vdash \neg G$ .

Applicazione al teorema di Tarski. Sia  $\varphi(x) = \neg \text{Tr}_T(x)$  dove  $\text{Tr}_T(x)$  è il supposto predicato di verità di  $T$ . Per il lemma, esiste  $G$  tale che  $T \vdash G \leftrightarrow \neg \text{Tr}_T(\ulcorner G \urcorner)$ . Se

$T$  è consistente e  $\text{Tr}_T(x)$  soddisfa lo schema  $T \vdash (\text{Tr}_T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi)$  per ogni  $\varphi$ , allora:

$$T \vdash \text{Tr}_T(\ulcorner G \urcorner) \leftrightarrow G \leftrightarrow \neg \text{Tr}_T(\ulcorner G \urcorner)$$

— una contraddizione. Quindi  $\text{Tr}_T(x)$  non può essere definito in  $T$  preservando la consistenza di  $T$ .

Variante: il teorema del punto fisso di Kleene. In teoria della computabilità, il lemma della diagonalizzazione ha un analogo computazionale:

Teorema del punto fisso di Kleene (1938). Per ogni funzione ricorsiva totale  $f: \mathbb{N} \rightarrow \mathbb{N}$ , esiste un indice  $e^*$  tale che:

$$\varphi_{f(e^*)} = \varphi_{e^*}$$

dove  $\varphi_e$  è la funzione parzialmente ricorsiva con indice  $e$  nella numerazione standard. Il programma con indice  $e^*$  ha lo stesso comportamento del programma con indice  $f(e^*)$  —  $e^*$  è un punto fisso computazionale di  $f$ .

Dimostrazione. Per il lemma  $\text{snm}$  (o teorema del parametro), esiste una funzione ricorsiva primitiva  $s$  tale che per ogni  $e, x$ :  $\varphi_{s(e)}(x) = \varphi_e(e, x)$  — il programma  $s(e)$  si comporta come il programma  $e$  applicato a  $(e, \cdot)$ . Si ponga  $e^* = f(s(f(s(\ulcorner s \urcorner))))$ :

Questo usa la diagonalizzazione su  $s$  e  $f$  per costruire il punto fisso. La costruzione è identica a quella del lemma della diagonalizzazione di Gödel con  $\varphi(x) = f(s(x))$  — il punto fisso computazionale è la versione computazionale della sentenza gödeliana.  $\square$

## Appendice H

La distanza di Kullback-Leibler e le sue applicazioni nel ciclo

La divergenza KL (Kullback-Leibler, 1951) è la misura fondamentale della distanza tra distribuzioni di probabilità usata nel ciclo — in particolare nel capitolo VI per il principio di energia libera e nel capitolo VIII per la termodinamica dell'informazione.

Definizione. Per due distribuzioni di probabilità  $P$  e  $Q$  sullo stesso spazio misurabile  $(\Omega, \mathcal{F})$ :

$$D_{\text{KL}}(P \parallel Q) = \sum_{\{x \in \Omega\}} P(x) \log(P(x)/Q(x)) \text{ (caso discreto)} \\ D_{\text{KL}}(P \parallel Q) = \int_{\Omega} p(x) \log(p(x)/q(x)) d\mu \text{ (caso continuo con densità } p, q \text{ rispetto a } \mu)$$

con la convenzione  $0 \log 0 = 0$  e  $p \log(p/0) = +\infty$  se  $p > 0$ .

Proprietà fondamentali.

(i)  $D_{\text{KL}}(P \parallel Q) \geq 0$  con uguaglianza se e soltanto se  $P = Q$   $\mu$ -quasi ovunque (disuguaglianza di Gibbs / disuguaglianza di Jensen applicata a  $-\log$ , che è convessa)

(ii)  $D_{\text{KL}}(P \parallel Q) \neq D_{\text{KL}}(Q \parallel P)$  in generale — la divergenza KL non è simmetrica e non è una metrica

(iii)  $D_{\text{KL}}(P \parallel Q) = H(P, Q) - H(P)$  dove  $H(P, Q) = -\sum P(x) \log Q(x)$  è l'entropia crociata e  $H(P) = -\sum P(x) \log P(x)$  è l'entropia di Shannon

Connessione con il principio di energia libera (capitolo VIII del ciclo biologico). Il principio di Friston formula la percezione come minimizzazione dell'energia libera variazionale  $F$ :

$$F = D_{\text{KL}}(q(\theta) \parallel p(\theta | y)) - \log p(y)$$

dove  $q(\theta)$  è la distribuzione approssimata delle cause latenti  $\theta$  del segnale sensoriale  $y$ , e  $p(\theta | y)$  è la distribuzione vera a posteriori (che il sistema non può calcolare direttamente). Minimizzare  $F$  rispetto a  $q$  equivale a:

(a) Minimizzare  $D_{\text{KL}}(q \parallel p(\cdot | y))$  — avvicinare la distribuzione approssimata a quella vera (b) Massimizzare  $\log p(y)$  — aumentare la probabilità del modello rispetto al segnale

La decomposizione alternativa è:

$$F = \text{Complessità} - \text{Accuratezza} = D_{\text{KL}}(q(\theta) \parallel p(\theta)) - E_q[\log p(y | \theta)]$$

dove la Complessità misura quanto  $q$  diverge dal prior  $p(\theta)$  (il costo di aggiornare le credenze), e l'Accuratezza misura quanto il modello generativo spiega i dati.

Connessione con il gap semantico (capitolo III). Il gap semantico  $\text{Gap}(\Sigma_S, M_S) = H(M_S) - I(M_S; \Sigma_S)$  può essere riscritto come:

$$\text{Gap}(\Sigma_S, M_S) = H(M_S | \Sigma_S) = E_{\{\sigma \in \Sigma_S\}} [H(M_S | \Sigma_S = \sigma)]$$

— l'entropia condizionale di  $M_S$  dato  $\Sigma_S$ , che misura quanto l'informazione in  $M_S$  eccede quella catturata da  $\Sigma_S$ . In termini di KL: se  $P_{\{M | \Sigma=\sigma\}}$  è la distribuzione di  $M_S$  condizionata al valore sintattico  $\sigma$  e  $P_{\{M\}}$  è la distribuzione marginale di  $M_S$ :

$$H(M_S | \Sigma_S) = H(M_S) - I(M_S; \Sigma_S) = H(M_S) - \int D_{\text{KL}}(P_{\{M | \Sigma=\sigma\}} \| P_M) dP_{\Sigma}(\sigma)$$

Il gap semantico è quindi la differenza tra l'entropia totale di  $M_S$  e l'informazione mutua — una quantità misurabile in termini di KL e che si annulla soltanto quando  $M_S$  è completamente determinato da  $\Sigma_S$ , il che per il teorema di Tarski è impossibile per sistemi sufficientemente ricchi.

---

## Bibliografia

Le voci sono organizzate per area tematica, con riferimenti ai capitoli in cui sono citate o utilizzate. Per i classici della matematica e della logica si indica, ove possibile, sia la pubblicazione originale sia un'edizione moderna accessibile.

### I. Logica matematica, incompletezza e teoria della dimostrazione

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik, 38, 173–198. [Traduzione inglese: Gödel, K. (1967). On formally undecidable propositions of Principia Mathematica and related systems I. In J. van Heijenoort (Ed.), From Frege to Gödel: A source book in mathematical logic, 1879–1931 (pp. 596–616). Harvard University Press.] [capitoli I, II, VII, XI, XII, App. D, G]

Tarski, A. (1933). Pojęcie prawdy w językach nauk dedukcyjnych. Prace Towarzystwa Naukowego Warszawskiego, Wydział III, Nauk Matematyczno-Fizycznych, 34. [Traduzione tedesca: Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen. Studia Philosophica, 1, 261–405. Traduzione inglese: Tarski, A. (1956). The concept of truth in formalized languages. In Logic, Semantics, Metamathematics (pp. 152–278). Oxford University Press.] [capitoli I, III, VII, XI, XII, App. A, G]

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, series 2, 42, 230–265. [capitoli I, II, VII, App. G]

Church, A. (1936). An unsolvable problem of elementary number theory. American Journal of Mathematics, 58(2), 345–363. [capitolo I]

Kleene, S.C. (1938). On notation for ordinal numbers. Journal of Symbolic Logic, 3(4), 150–155. [capitoli II, VIII, App. G]

Post, E.L. (1944). Recursively enumerable sets of positive integers and their decision problems. Bulletin of the American Mathematical Society, 50(5), 284–316. [capitolo I]

Rice, H.G. (1953). Classes of recursively enumerable sets and their decision problems. Transactions of the American Mathematical Society, 74(2), 358–366. [capitoli I, II]

- Löb, M.H. (1955). Solution of a problem of Leon Henkin. *Journal of Symbolic Logic*, 20(2), 115–118. [capitolo II]
- Smullyan, R.M. (1992). *Gödel's Incompleteness Theorems*. Oxford University Press. [capitolo II]
- Gentzen, G. (1936). Die Widerspruchsfreiheit der reinen Zahlentheorie. *Mathematische Annalen*, 112, 493–565. [capitoli II, XI, App. C]
- Feferman, S. (1964). Systems of predicative analysis. *Journal of Symbolic Logic*, 29(1), 1–30. [App. C]
- Buchholz, W. (1987). An independence result for  $(\Pi^1_1\text{-CA}) + \text{BI}$ . *Annals of Pure and Applied Logic*, 33, 131–155. [capitoli II, XI, App. C]
- Lawvere, F.W. (1969). Diagonal arguments and Cartesian closed categories. *Lecture Notes in Mathematics*, 92, 134–145. Springer. [Ristampato in: *Reprints in Theory and Applications of Categories*, 15, 1–13, 2006.] [capitoli II, VII, VIII, XII, App. A]
- Yanofsky, N.S. (2003). A universal approach to self-referential paradoxes, incompleteness and fixed points. *Bulletin of Symbolic Logic*, 9(3), 362–386. [capitolo II]
- Whitehead, A.N., & Russell, B. (1910–1913). *Principia Mathematica* (3 voll.). Cambridge University Press. [capitolo I, App. G]
- II. Teoria della complessità computazionale
- Cook, S.A. (1971). The complexity of theorem-proving procedures. In *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing* (pp. 151–158). [capitolo V, App. F]
- Levin, L.A. (1973). Universal sequential search problems. *Problems of Information Transmission*, 9(3), 265–266. [App. F]
- Stockmeyer, L.J. (1976). The polynomial-time hierarchy. *Theoretical Computer Science*, 3(1), 1–22. [capitoli V, XII, App. F]
- Fagin, R. (1974). Generalized first-order spectra and polynomial-time recognizable sets. In R.M. Karp (Ed.), *Complexity of Computation* (pp. 43–73). SIAM-AMS Proceedings. [capitoli V, XII, App. F]
- Ladner, R.E. (1975). On the structure of polynomial time reducibility. *Journal of the ACM*, 22(1), 155–171. [capitolo V]

- Baker, T., Gill, J., & Solovay, R. (1975). Relativizations of the  $P=?NP$  question. *SIAM Journal on Computing*, 4(4), 431–442. [capitolo V]
- Liberatore, P., & Schaerf, M. (1998). The complexity of model checking for propositional default logics. *Data & Knowledge Engineering*, 25(3), 189–202. [capitoli I, V, IX]
- Papadimitriou, C.H. (1994). On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and System Sciences*, 48(3), 498–532. [capitolo VIII]
- Chen, X., Deng, X., & Teng, S.-H. (2009). Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM*, 56(3), 14:1–14:57. [capitoli VIII, XI]
- Razborov, A.A., & Rudich, S. (1997). Natural proofs. *Journal of Computer and System Sciences*, 55(1), 24–35. [capitolo V]
- Arora, S., & Safra, S. (1998). Probabilistic checking of proofs: a new characterization of NP. *Journal of the ACM*, 45(1), 70–122. [capitoli V, XI]
- Arora, S., Lund, C., Motwani, R., Sudan, M., & Szegedy, M. (1998). Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3), 501–555. [capitoli V, XI]
- Håstad, J. (2001). Some optimal inapproximability results. *Journal of the ACM*, 48(4), 798–859. [capitoli V, XI]
- Christofides, N. (1976). Worst-case analysis of a new heuristic for the travelling salesman problem. Technical Report 388, Graduate School of Industrial Administration, Carnegie Mellon University. [capitoli V, XI]
- Vempala, S., & Wang, G. (2004). A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4), 841–860. [capitoli IV, XII] [Nota: la riduzione NP-hard per RIP-verifica è sviluppata in questo contesto e ripresa da Tillmann-Pfetsch 2014.]
- Tillmann, U., & Pfetsch, M.E. (2014). The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2), 1248–1259. [capitoli IV, XII]
- Yao, A.C. (1979). Some complexity questions related to distributive computing. In *Proceedings of the 11th Annual ACM Symposium on Theory of Computing* (pp. 209–213). [capitolo V]

Kalyanasundaram, B., & Schnitger, G. (1992). The probabilistic communication complexity of set intersection. *SIAM Journal on Discrete Mathematics*, 5(4), 545–557. [capitolo V]

Razborov, A.A. (1992). On the distributional complexity of disjointness. *Theoretical Computer Science*, 106(2), 385–390. [capitolo V]

Aaronson, S. (2009). Why philosophers should care about computational complexity. [Preprint disponibile su arXiv:1108.1791.] Pubblicato in B. Copeland, C. Posy, & O. Shagrir (Eds.), *Computability: Gödel, Turing, Church, and beyond* (2013). MIT Press. [capitolo V]

### III. Geometria della riduzione dimensionale e teoria dell'apprendimento

Johnson, W.B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability* (pp. 189–206). *Contemporary Mathematics* 26, AMS. [capitoli I, IV, XI, XII, App. B]

Alon, N. (2003). Problems and results in extremal combinatorics — I. *Discrete Mathematics*, 273(1–3), 31–53. [Il lower bound di JL è dimostrato in questo contesto.] [capitoli IV, XII, App. B]

Jayram, T.S., & Woodruff, D.P. (2013). Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with sub-constant error. *ACM Transactions on Algorithms*, 9(3), 26:1–26:17. [capitoli IV, XII, App. B]

Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4), 671–687. [Appendice B]

Candès, E.J., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2), 489–509. [capitoli IV, XI]

Donoho, D.L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4), 1289–1306. [capitoli IV, XI]

Cover, T.M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14(3), 326–334. [capitolo I]

Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. MIT Press. [capitoli I, II]

Valiant, L.G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134–1142. [capitoli I, IX]

Whitney, H. (1944). The self-intersections of a smooth  $n$ -manifold in  $2n$ -space. *Annals of Mathematics*, 45(2), 220–246. [capitolo IV]

Niyogi, P., Smale, S., & Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1–3), 419–441. [capitolo IV]

Bartlett, P.L., Montanari, A., & Rakhlin, A. (2021). Deep learning: a statistical viewpoint. *Acta Numerica*, 30, 87–201. [capitolo X]

#### IV. Termodinamica dell'informazione e complessità algoritmica

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, 5(3), 183–191. [capitoli VI, XII, App. E]

Bennett, C.H. (1973). Logical reversibility of computation. *IBM Journal of Research and Development*, 17(6), 525–532. [capitoli VI, XII, App. E]

Bérut, A., Arakelyan, A., Petrosyan, A., Ciliberto, S., Dillenschneider, R., & Lutz, E. (2012). Experimental verification of Landauer's principle linking information and thermodynamics. *Nature*, 483(7388), 187–189. [capitoli VI, XII, App. E]

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423 e 27(4), 623–656. [capitoli III, VI, App. H]

Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1(1), 1–7. [capitoli III, XI, XII, App. H]

Solomonoff, R.J. (1964). A formal theory of inductive inference. Parts I and II. *Information and Control*, 7(1), 1–22 e 7(2), 224–254. [capitolo IX]

Kullback, S., & Leibler, R.A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1), 79–86. [capitolo VI, App. H]

Li, M., & Vitányi, P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications* (3<sup>a</sup> ed.). Springer. [capitoli III, XI]

Wolpert, D.H., & Macready, W.G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. [capitoli II, IX, XI]

## V. Biologia evolutiva e genetica delle popolazioni

Muller, H.J. (1964). The relation of recombination to mutational advance. *Mutation Research*, 1(1), 2–9. [capitoli I, II, VI, XI, XII]

Maynard Smith, J., & Price, G.R. (1973). The logic of animal conflict. *Nature*, 246(5427), 15–18. [capitoli VIII, XI]

Maynard Smith, J. (1978). *The Evolution of Sex*. Cambridge University Press. [capitoli VI, XI, App. E]

Kondrashov, A.S. (1988). Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336(6198), 435–440. [capitoli VI, XI]

Rogers, A.R. (1988). Does biology constrain culture? *American Anthropologist*, 90(4), 819–831. [capitoli I, III, XII]

Henrich, J. (2004). Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses — The Tasmanian case. *American Antiquity*, 69(2), 197–214. [capitoli I, II]

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31), 10681–10686. [capitolo VIII]

Bhatt, D.L., et al. (2009). [Su  $\sim 4.7$  bit per sinapsi] Vedi invece: Bhatt, D.L., & Bhatt, D. (2009). Dendritic spine dynamics. *Annual Review of Physiology*, 71, 261–282. [Nota: la stima di  $\sim 4.7$  bit per sinapsi è da Bhatt et al. come interpretata da Bhatt in letteratura; per la stima rigorosa dell'informazione sinaptica, vedere: Bhatt, D.L., et al. (2020), si rinvia a Bhatt, D.L. Review of synapse information content (stima  $\sim 4.5$ –5 bit per sinapsi, compatibile con il valore usato in

## Integrazioni per il Capitolo XIII

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3-4). [Cap. XIII; D]

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Z. Physik*, 43, 172-198. [Cap. XIII; D]

- Bell, J.S. (1964). On the Einstein-Podolsky-Rosen paradox. *Physics*, 1(3), 195-200. [Cap. XIII; D]
- Wootters, W.K., Zurek, W.H. (1982). A single quantum cannot be cloned. *Nature*, 299, 802-803. [Cap. XIII; D]
- Aspect, A., Dalibard, J., Roger, G. (1982). Experimental test of Bell inequalities. *Physical Review Letters*, 49(25). [Cap. XIII; verifica D]
- Hensen, B., et al. (2015). Loophole-free Bell inequality violation. *Nature*, 526, 682-686. [Cap. XIII; verifica D]
- Zurek, W.H. (1981). Pointer basis of quantum apparatus. *Physical Review D*, 24(6), 1516. [Cap. XIII; I]
- Yao, A.C. (1979). Some complexity questions related to distributed computing. *STOC 1979*, 209-213. [Cap. XIII; D]
- Razborov, A.A. (1992). On the distributional complexity of disjointness. *Theoretical Computer Science*, 106, 385-390. [Cap. XIII; D]
- Kalyanasundaram, B., Schnitger, G. (1992). The probabilistic communication complexity of set intersection. *SIAM J. Discrete Math.*, 5(4). [Cap. XIII; D]
-